'HE AMERICAN ECONOMIC REVIEW

June 1979

VOLUME 69, NUMBER 3

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H Borts, Managing Editor, Box Q, Brown University, Providence, R I 02912 Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript A submission fee must accompany each manuscript $15 for members; $30 for nonmembers *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor

Articles

Shorter Papers

Notes

THE AMERICAN ECONOMIC REVIEW

September 1979

VOLUME 69, NUMBER 4

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members; $30 for nonmembers *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

Articles

Shorter Papers

THE AMERICAN ECONOMIC REVIEW

December 1979

VOLUME 69, NUMBER 5

• Manuscripts and editorial correspondence relating to the regular quarterly issue of this *REVIEW* and the *Papers and Proceedings* should be addressed to George H. Borts, Managing Editor, Box Q, Brown University, Providence, R.I. 02912. Manuscripts should be submitted in duplicate and in acceptable form and should be no longer than 50 pages of double-spaced typescript. A submission fee must accompany each manuscript: $15 for members, $30 for nonmembers. *Style Instructions* for guidance in preparing manuscripts will be provided upon request to the editor.

• No responsibility for the views expressed by authors in this *REVIEW* is assumed by the editors or the publishers, The American Economic Association.

Articles

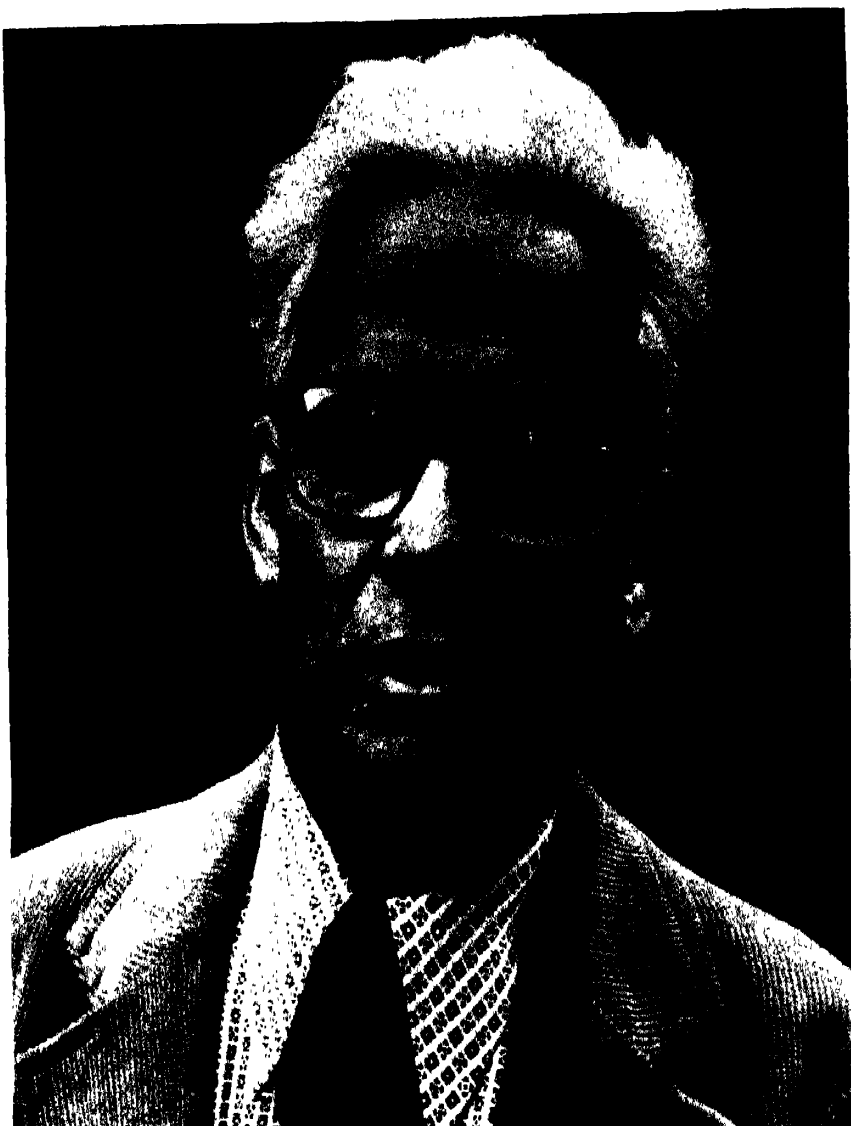
Shorter Papers

RICHARD A. MUSGRAVE

DISTINGUISHED FELLOW

1978

Richard A. Musgrave has been the dean of public finance economists since the end of World War II. Much of what we now know as public finance bears the imprint of his insights and analytical innovations. He originated the division of budget policy into the three branches of Allocation, Distribution, and Stabilization. He contributed to the theory of public goods, and developed the concept of merit wants. He was the first to analyze the incidence of taxation from both the uses and sources side of income, and pioneered the development of estimates of the distribution of tax burdens along these lines. He helped refine, and developed new measures of, built-in flexibility of government budgets. He participated in the development of the theory of risk taking when a tax is imposed on investment income.

But Richard Musgrave has not been an ivory-tower economist. He has made major contributions to the measurement of the incidence and economic effects of expenditure and tax policies. He has been a policy advisor to the United States and numerous other governments. He has led tax missions to several countries and his recommendations have had a lasting impact on the structure of taxation in these countries.

Richard A. Musgrave

Industry Performance Gradient Indexes

By ROBERT E. DANSBY AND ROBERT D. WILLIG*

This paper presents a theory of indexes which measure the rate of potential improvement in the welfare performance of an industry. These indexes indicate the magnitude of gross social gains achievable from appropriate governmental intervention (for example, antitrust, regulatory and deregulatory actions, or threats thereof). The indexes are local measures which can be calculated from data pertaining to the current industry structure (i.e., market shares and demand elasticities). Surprisingly, the indexes reduce to simple transformations of standard indexes of market concentration[1] and monopoly power (namely the m-firm concentration ratio, the Herfindahl index, and the Lerner index) given familiar sets of assumptions on firm behavior (respectively: collusive price-leadership, quantity Cournot, and pure monopoly).

Since different modes of firms' conduct lead to different indexes, the choice among concentration index formulae should be based on an assessment of the behavior of the industry's firms. We find that the potential improvement in welfare performance is as sensitive to mode of conduct and other industry data as it is to the observed market shares. Consequently, our analysis provides a quantification of the idea that concentration per se does not necessarily warrant governmental intervention. Our theory at once provides a general index concept, new rigorously based practical indexes, a conceptual framework for the interpretation of standard indexes, and insights into appropriate criteria for governmental intervention.

A rational appraisal of the desirability of a governmental action towards an industry can be phrased as a comparison of the benefits and the costs of the intervention. Each of the many possible governmental actions can conceptually be associated with the vectors q^0 and q of the outputs of the firms in the industry, before and after the intervention, respectively. The gross benefits of each action may be expressed as $W(q) - W(q^0)$, where $W(\cdot)$ is the sum of consumers' and producers' surpluses. While received theory does guide the specification of the social objective function, little can be said at this level of generality about the social cost of governmental action which moves industry outputs from q^0 to q. Even so, it is useful to examine the benefit side of the rational calculus of intervention.

It appears that the government regards an industry with high values of the standard concentration indexes as a prime candidate for intervention.[2] Thus, using the cost-benefit vocabulary, the prevailing view seems to be that the concentration indexes are strongly positively correlated with $W(q) - W(q^0)$, where q is the result of appropriate corrective action.

In this paper we synthesize the rigorous cost-benefit and the practical index number approaches to the identification of industries where the government's intervention efforts will be well placed. Our aim is to develop tools capable of assessing $W(q) - W(q^0)$. Yet, to ensure that the tools are practical ones, we accept constraints implicit in the index number methodology and confine ourselves to the use of information on only the *current* situation of the industry. Consequently, we focus on the rate of change of $W(\cdot)$ at q^0; that is, on the current sensitivity of social welfare

*American Telephone and Telegraph Company and Princeton University, respectively. This paper was written while we were employed by Bell Laboratories and is partly based on Dansby's doctoral dissertation. We are grateful to W. J. Baumol, A. Weiss, and S. Winter for extremely helpful comments and discussions.

[1]The measurement of industrial concentration is discussed by Morris Adelman, John Blair, and Russell Parker. The data used in these measurements typically come from Bureau of the Census or Federal Trade Commission sources. See J. E. Morton.

[2]Although economists debate the relative merits of various concentration indexes (see Eugene Singer or James Delaney), the government unabashedly uses these indexes to guide intervention activities (see F. M. Scherer).

to adjustments in the outputs of firms in the industry.

The complete specification of the local rate of change of $W(\cdot)$ is given by the gradient of W at q^0, the vector of partial derivatives, $W_i(q^0)$, of W with respect to the output of each firm i in the industry. The first-order estimate of $W(q) - W(q^0)$ is

$$\sum_i W_i(q^0)(q_i - q_i^0) \equiv \nabla W \cdot \Delta q$$

Further, for concave $W(\cdot)$, $\nabla W \cdot \Delta q$ gives an upper bound on $W(q) - W(q^0)$ even for global changes in quantities. Then, a *necessary condition* for an intervention activity to be socially desirable is that $\nabla W \cdot \Delta q$ exceed the social cost. Hence, for a given set of output changes each having the sign of $W_i(q^0)$, the larger in absolute value are the components of the gradient vector, the larger is the first-order change in welfare, and the more likely is the associated government action to fulfill this necessary condition for desirability. Note, however, that $\nabla W \cdot \Delta q$ depends on the specific direction of industry output changes. While this may be the appropriate approach for the evaluation of a particular intervention action, a practical characterization of *industry* performance sensitivity requires a scalar index which is independent of policy detail.

For this reason, we define the industry performance gradient index as $\phi = \nabla W \cdot \Delta q^*$ where the weights Δq^* are the components of the unit vector which points in the direction of steepest ascent of the objective function $W(\cdot)$. Thus, the new index ϕ measures the sensitivity of social welfare to the locally best changes in the outputs of the firms in an industry.

In Section I we relate the industry performance gradient index to the firms' proportional deviations between prices and marginal costs. We derive the fundamental properties of the index, and show how it can be interpreted as an indicator of the gross benefits from intervention. In Section II, we show that for a homogeneous product industry, the gradient vector can be written in terms of current industry price, price elasticity of demand, behavioral mode, and market shares. This enables us, in Section III, to study the influences of market shares and behavioral modes on the industry performance gradient index. The index is shown to reduce, in various behavioral scenarios, to various standard indexes of concentration and monopoly power.

In Section IV we develop different industry performance gradient indexes based on different metrics over the outputs of the industry's firms. These indexes lead to new practical measures of concentration for each mode of firm behavior. We argue that the most appropriate metric is the one most closely related to the costs of the governmental actions required to effect the movements. The various metrics investigated represent intervention cost functions with different returns to scale and with different sensitivities to firm size. Thus, the desirability of intervention can be crucially dependent on the cost characteristics of such action.

I. The Industry Performance Gradient Index and Its Properties

We assume a fixed population of n firms in the industry.[3] The ith firm produces output q_i, faces the differentiable inverse demand function $P_i(q_1, \ldots, q_n)$, has a differentiable production cost function $C_i(q_i)$, and earns profits $\pi_i = q_i P_i(q) - C_i(q_i)$. The function $W(q_1, \ldots, q_n)$ is continuously twice-differentiable with first derivatives that are proportional to those of the sum of $\Sigma_{i=1}^n \pi_i$ and Marshallian consumer's surplus. Further, we remove the proportionality constant and scale W in dollars.[4] Thus, in our partial equilibrium framework,

[3]Thus our analysis ignores considerations of entry and exit. However, the case of actions which split a firm can be handled by comparing the old output of the firm with the total output of its new fragments

[4]If W were taken to be the sum of producers' and consumers' surpluses, we would have to assume no income effects on each consumer's demands for the goods in order to make W well defined. We can avoid this restriction by taking W to be a Bergsonian welfare function of a policymaker who is indifferent to interpersonal transfers of nominal income. Then the derivatives of W are proportional to those of total surplus. However with this specification, W is not in money units and cannot be compared with the money costs of governmen

$$(1) \qquad W_i(q) \equiv \frac{\partial W(q)}{\partial q_i} = P_i(q) - C_i'(q_i)$$

where $C_i'(q_i)$ is the marginal cost of the ith firm when it produces q_i.

Figure 1 displays level curves, with $W^0 < W^1 < W^2$, of the welfare function of a given industry consisting of two firms whose outputs are q_1 and q_2. Given the current output vector represented by A in Figure 1, welfare can be increased to the level W^1 by moving the outputs of the firms to the points B, C, or D, for example. It is clear that the rate of increase in welfare is larger when output is moved from A to D than when it is moved from A to C. In the former case output is increased by fewer units than in the latter case, to effect the same increase in welfare. This can be an important distinction when the social costs of corrective action increase with the size of the output adjustment.

We define the industry performance gradient index to be the maximum rate of increase in social welfare at the current output point. At point A in Figure 1 the locally best direction of change which achieves this maximal rate is represented by the solid arrow pointing toward D. Of course, the maximum rate of increase in social welfare depends on the vector of current outputs. For example, supposing that $W^2 - W^1 = W^1 - W^0$, the maximal rate is larger at A than at D. Thus, for any small adjustment in outputs from D, there is an equal sized adjustment from A which accomplishes a greater improvement in welfare (see Proposition 5, below). In this way, the industry performance gradient index gives an intuitively meaningful ranking of industry positions.

The choice of a metric to measure the magnitude of output adjustments is a delicate and important matter. The obvious candidate is the usual geometric (Euclidean) metric, $[\Sigma(\Delta q_i)^2]^{1/2}$, which treats all $\Delta q_i = q_i - q_i^0$

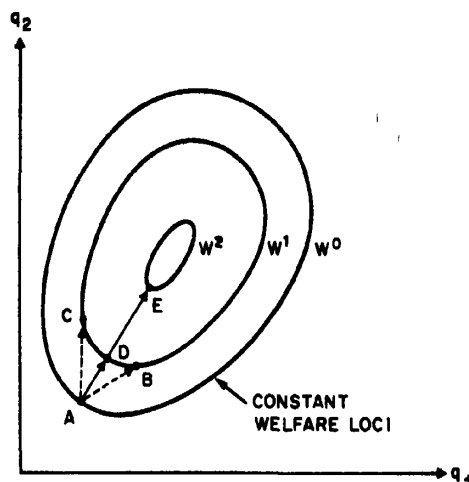



FIGURE 1

symmetrically. However, in the case of heterogeneous products, q_i and q_j are quantities of different commodities, and are possibly measured in different units. Hence, scaling factors must be applied to make the values commensurable. For goods produced in the same industry, it seems reasonable to accept the base prices as scaling factors. Thus, in this section, we construct an index from the foundation of the base price weighted Euclidean distance:

$$(2) \qquad \rho(\Delta q, P^0) = [\Sigma(P_i^0 \Delta q_i)^2]^{1/2}$$

where Δq and P^0 denote vectors. The units of this distance measure are those of money.

Our prime concern is to investigate the sensitivity of the level of welfare to the permitted magnitude of optimally chosen output adjustments. Specifically, we solve for the $q^*(t)$ which maximizes W within the "hypersphere," $\rho(\Delta q, P^0) \leq t$, with radius t and center q^0. Thus, $W(q^*(t)) - W(q^0)$ is the largest increase in welfare achievable by output adjustments confined to the distance t from q^0. The corresponding best direction of change in the outputs of the industry's firms is proportional to $(q^*(t) - q^0)$. The average rate of change of welfare with respect to the distance measure, when outputs are adjusted in the best direction, is $(W(q^*(t)) - W(q^0)) \div t$. The industry performance

tal intervention. This problem can be obviated by considering those costs deducted from individuals' incomes, inside the Bergsonian function. To leave the text free of such verbiage, we shall speak of W as a welfare function scaled in dollars.

gradient index ϕ is defined as the best instantaneous rate of change of W, at the current industry position; that is,

$$(3) \qquad \phi = \lim_{t \to 0} [W(q^*(t)) - W(q^0)] \div t$$

THEOREM 1:[5] *For the metric defined in* (2), *the industry performance gradient index at q^0 is*

$$(4) \qquad \phi = [\Sigma((P_i^0 - C_i'(q_i^0))/P_i^0)^2]^{1/2}$$

Further, the associated best direction of adjustment in the firms' base price weighted outputs is

$$(5) \quad d(P_i^0 q_i^*(t))/dt|_{t=0} = [P_i^0 - C_i'(q_i^0)]/\phi P_i^0$$

These results tell a sensible story. The local sensitivity of welfare to the optimal quantity adjustments is directly related to the absolute size of the percentage deviations of prices from marginal costs. If, for good i, price exceeds marginal cost, then the best local direction of adjustment has q_i increase. Moreover, across goods, the larger percentage deviations command the larger price weighted quantity adjustments. However, if all industry prices were equal to the associated marginal costs, then q^0 would be a local maximum of W, the best local change in quantities would be no change at all, and $\phi(q^0) = 0$. Inspection of (4) leads to

PROPOSITION 1: $\phi(q^0) = 0$ *if and only if* $P_i(q^0) = C_i'(q_i^0)$ *for all i. Further, zero is the minimum value of* ϕ.

If $W(\cdot)$ is a strictly concave function, then a q at which $\phi = 0$ is the unique global welfare optimum. For any other q, W is smaller, while ϕ is larger. One might thus be lead to believe that ϕ can be used as an inverse indicator of welfare performance. In fact, we do have this result:

PROPOSITION 2:[6] *If output adjustments are continually made in the best local direction given by* (5), *then, while W increases,* ϕ *will fall monotonically if* $W(\cdot)$ *is strictly concave.*

Figure 2 pictures a strictly concave $W(q_1, q_2)$ surface. The quantity vector associated with point B, q^B, is drawn to be the outcome of a series of best direction adjustments starting from q^A. As such, $W(q^B) > W(q^A)$. Also, as required by Proposition 2, a comparison of the steepest slopes of the W hill at the two points indicates that $\phi(q^B) < \phi(q^A)$. Thus, for quantity vectors on a single best adjustment path, ϕ and W are inversely related.

However, over pairs of industry output vectors without the special relationship posited in Proposition 2, ϕ cannot be expected to correctly rank levels of W. The heights of the points A and C are drawn to show that $W(q^A) < W(q^C)$. Yet, a comparison of the steepest slopes of the hill at these two points indicates that $\phi(q^A) < \phi(q^C)$. Thus, while q^C is a better industry position than q^A, the potential welfare gain from a given small size adjustment in outputs is larger at q^C than at q^A. Let us emphasize this distinction.

PROPOSITION 3:[7] *Among industry positions and across different industries,* ϕ *does*

[5]PROOF: The Lagrangian for the maximization of $W(q)$ subject to $\rho(\Delta q, P^0) \leq t$ can be written $L = W(q) + \lambda[t^2 - \Sigma(\Delta q_i P_i^0)^2]$. The first-order conditions, necessarily satisfied at the solution $q^*(t)$, λ, are: (a) $W_i(q^*(t)) = 2\lambda P_i^{02}(q_i^*(t) - q_i^0)$ and (b) $\lambda[t^2 - \Sigma(\Delta q_i P_i^0)^2] = 0$, $\lambda \geq 0$, and $\rho(\Delta q, P^0) \leq t$. By the Envelope Theorem (see Paul Samuelson or Eugene Silberberg), $dW(q^*(t))/dt = \partial L/\partial t = 2\lambda t$. Squaring (a) and dividing through by P_i^{02} for each i, summing over i, and using (b) yields $2\lambda t = [\Sigma W_i\ (q^*(t))^2/P_i^{02}]^{1/2}$. As $t \to 0$, $q^*(t) \to q^0$ and, by continuity, $W_i(q^*(t)) \to W_i(q^0)$. Then

$$\phi = \lim_{t \to 0} dW(q^*(t))/dt = \lim_{t \to 0} 2\lambda t$$

and (1) yields (4). Equation (5) follows from using (a) to solve for the limit, as $t \to 0$, of $[q_i^*(t) - q_i^0]/t$.

[6]PROOF: Let u be a nonnegative real variable parameterizing the path of q. At every u, q is adjusted in the best local direction given by (5). Thus $dq_j/du = W_j(q)/P_j^{02}\ \phi(q)$. Through q, $\phi(q)$ is a function of u. Differentiating (4) totally, $d\phi[q(u)]/du = \Sigma_i\Sigma_j(W_i W_{ij} W_j/P_i^{02}P_j^{02})/\phi^2$. This quadratic form is negative due to the assumed strict concavity of W and so $d\phi[q(u)]/du < 0$.

[7]In Section III, we study cases in which ϕ reduces to simple transformations of standard concentration indexes. These *are* often interpreted as indicators of industry performance.

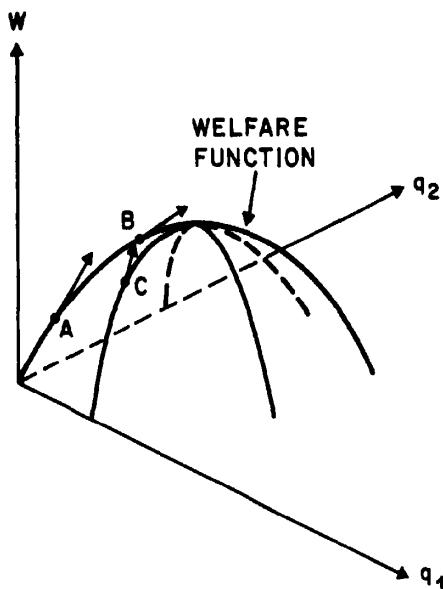



FIGURE 2

rank the potential gains from local quantity adjustments of given size. However, ϕ does not rank the levels of performance of different industries nor different positions of one industry.

To assure that our measure of potential industry welfare improvement can, like indexes of market concentration, be calculated from information on only the current state of the industry, we have defined it as a strictly local rate of change. Nonetheless, implications can be drawn from ϕ about welfare gains resulting from global adjustments in industry outputs.

PROPOSITION 4:[8] *If $W(\cdot)$ is concave then*

(6) $$W(q) - W(q^0) \leq \phi(q^0)\rho(\Delta q, P^0)$$

Thus, ϕ calculated at the current industry position can be used to obtain an upper bound on the welfare gains from arbitrary movements in the outputs of the industry's firms.

[8]PROOF: $W(q) - W(q^0) \leq \Sigma\,(W_i(q^0)/P_i^0)\,(P_i^0\,(q_i - q_i^0)) \leq (\Sigma\,(W_i(q^0)/P_i^0)^2)^{1/2} \times (\Sigma(P_i^0\,[q_i - q_i^0])^2)^{1/2} = \phi(q^0)\rho(\Delta q, P^0)$. The first inequality is a standard property of concave functions and the second is the Cauchy-Schwartz inequality.

The inequality in (6) results from the rate of change of W in any direction in quantity space being compared with the maximal rate of change.

Proposition 4 may be quite useful for discouraging socially wasteful intervention. If the costs of the government action needed to move q^0 to q are projected to be greater than $\phi(q^0)\,\rho(\Delta q, P^0)$, then (6) can be used to infer that the costs outweigh the potential benefits of the intervention. A necessary condition for the action to be beneficial is that $\phi(q^0)\cdot\rho(\Delta q, P^0)$ exceed the social costs. Attention to this rule can certainly help to structure public debate over proposed intervention policy.

The industry performance gradient index may also be useful in making comparisons of potential welfare gains among different industries or among different positions of a particular industry. Let the presuperscripts A and B denote values pertaining to two different industries, or to two different industry positions.

PROPOSITION 5:[9] *If ${}^A\phi({}^Aq^0) > {}^B\phi({}^Bq^0)$, then, for some $t^* > 0$, for any ${}^Bq$ with $\rho(\Delta^Bq, {}^BP^0) \leq t^*$, there is a ${}^Aq$ with $\rho(\Delta^Aq, {}^AP^0) = \rho(\Delta^Bq, {}^BP^0)$ and ${}^AW({}^Aq) - {}^AW({}^Aq^0) > {}^BW({}^Bq) - {}^BW({}^Bq^0)$.*

Suppose that the market performance gradient index is larger at the current position of industry A than it is at the current position of industry B. The proposition asserts that there is a distance t^* such that the welfare improvement from any adjustment of industry B's outputs within that distance can be exceeded by some equal sized adjustment, measured by ρ, in the output of industry A. It is in this sense that, for the same cost, industry A is a better candidate for intervention than B, if ${}^A\phi > {}^B\phi$.

[9]PROOF: Consider $\psi(t) \equiv ({}^AW({}^Aq^*(t)) - {}^AW({}^Aq^0)) - ({}^BW({}^Bq^*(t)) - {}^BW({}^Bq^0))$. By the definition of ϕ, at $t = 0$, $d\psi(t)/dt = {}^A\phi({}^Aq^0) - {}^B\phi({}^Bq^0)$, which is positive by hypothesis. Then, by a fundamental property of derivatives, there is a $t^* > 0$ such that $\psi(t) > \psi(0)$ for $0 < t \leq t^*$. Note that $\psi(0) = 0$. Let $\rho(\Delta^Bq, {}^BP^0) = t \leq t^*$. By the definition of ${}^Bq^*(t)$, ${}^BW({}^Bq) - {}^BW({}^Bq^0) \leq {}^BW({}^Bq^*(t)) - {}^BW({}^Bq^0)$, which is itself strictly less than ${}^AW[{}^Aq^*(t)] - {}^AW({}^Aq^0)$ because $\psi(t) > 0$.

II. The Welfare Gradient for Firms in a Homogeneous Product Industry

In this section we uncover relationships between the welfare gradient and homogeneous product industry data that are more easily obtained and interpreted than are the firms' marginal costs that are crucial in (4).

In a homogeneous product industry there is but one market price, and it can be viewed as a function, $P(Q)$, of the total industry output, $Q = \Sigma_{i=1}^{n} q_i$. The ith firm's profit is $\pi_i = q_i P(Q) - C_i(q_i)$. We assume that the ith firm sets its output to maximize its profit, subject to its perception of the effects its output adjustments have on the outputs of other firms, dq_j/dq_i. The first-order condition for the profit-maximizing q_i is

$$(7) \quad P(Q) + q_i\Big(1 + \sum_{j \neq i} \frac{dq_j}{dq_i}\Big) P'(Q) - C_i'(q_i) = 0$$

Letting $\alpha_i \equiv \Sigma_{j \neq i}\, dq_j/dq_i$ denote the total perceived effect of adjustments in q_i on all other firms' outputs, $\epsilon \equiv -P/P'Q$ denote the price elasticity of industry demand and $s_i \equiv q_i/Q$ denote the market share of the ith firm, (7) becomes[10]

$$(8) \qquad P - C_i'(q_i) = \frac{P}{\epsilon} s_i(1 + \alpha_i)$$

Thus, in view of (1), under our current assumptions the industry performance gradient is the vector whose ith component is given by (8).

The larger is $W_i(q^0)$ in absolute value, the more sensitive is industry performance to given changes in q_i. When comparing $W_i(q^0)$ across firms it is meaningful to infer from (8) that a larger observed price means a larger value of W_i, other things equal. This is simply because consumers impute the value of PdQ to local increases in industry output.

The elasticity of market price with respect to the output of the ith firm is s_i/ϵ. Hence, the larger the firm's market share, and the smaller the price elasticity of industry demand, the larger is the price decrease that results from a given proportional increase in q_i, and the larger is W_i. This finding supports previous arguments that the elasticity of industry demand has an influence on the desirability of government intervention. (See, for example, A. C. Johnson and Peter Helmberger.) Further, it supports the common notion that firms with larger market shares are the most appropriate candidates for corrective actions that would stimulate greater output from the firm's productive facilities.[11]

Another observation is that α_i, the perceived industry output reaction to a change in q_i, is an important determinant of W_i.[12] The more positive is α_i, the larger is the elasticity of price with respect to q_i perceived by firm i, and the larger is the deviation between marginal cost and price.

For given size quantity adjustments, these results provide: (a) quantifiable criteria for judging which firms can contribute the largest welfare gains; and (b) a method to compute, via $\nabla W \cdot \Delta q$, an estimate of or an upper bound on the welfare increment. However, in the absence of a particular policy proposal, or a set of predicted output changes, the industry performance gradient vector does not provide a practical measure of the sensitivity of *industry* performance. For this role, a scalar industry-specific indicator is required.

III. The Performance Gradient Index for Homogeneous Product Industries

Combination of the generic index formula (4) with the characterization of firm behavior given by (8) yields the following form of the industry performance gradient index:

$$(9) \qquad \phi = \frac{1}{\epsilon}\,[\Sigma s_i^2(1 + \alpha_i)^2]^{1/2}$$

The index ϕ is inversely related to ϵ and is increasing in positive α_i's, as was W_i. Unlike

[10] This relationship was derived and utilized in a different context by Keith Cowling and M. Waterson.

[11] Sidney Winter has suggested to us that forced fission of a firm can be viewed in this way, inasmuch as it transforms a single centralized locus of control into several competing units.

[12] An index proposed by Kurt Rothschild attempts to measure the effects of such interfirm reactions.

W_i, however, ϕ is not a direct function of the market price because its underlying metric is itself scaled in units of value. While it was clear that W_i increases with s_i, the behavior of ϕ with respect to the vector of market shares, $s = (s_1, \ldots, s_n)$, is more complex since, if one share increases, another must decrease to maintain $\Sigma s_i = 1$.

It can be seen from (9) that ϕ is an increasing quasi-convex function in s.[13] Thus, as pictured in Figure 3, its level curves are concave to the origin, and ϕ increases from "inner" to "outer" curves. Of course, the only relevant market share vectors are those on the nonnegative unit simplex, $\Sigma s_i = 1$. If s' and s'' are such vectors which, other things equal, make ϕ the same, then ϕ is smaller for any convex combinations, $s''' = \theta s' + (1 - \theta)s''$, with $1 > \theta > 0$. If the coefficients of conjectural variation, α_i, are the same for all firms, or if $(1 + \alpha_i)^2$ increases with market share, then a shift in share (*ceteris paribus*) from a smaller firm to a larger one will increase ϕ, and ϕ increases with the inequality of the market shares.[14] This is a property that indexes of market concentration are often required to exhibit (see Tibor Scitovsky).

In fact, the most widely utilized indexes of market concentration and monopoly power emerge as simple transformations of special cases of ϕ. For a profit-maximizing monopolist, the term in brackets in (9) equals 1 and $1/\epsilon = \{P - C'(q)\}/P$.

PROPOSITION 6: *In a single product, monopolized industry, $\phi = (P - C')/P$, which is the Lerner Index of monopoly power. Hence, the industry performance gradient index can be viewed as a natural generalization of the Lerner Index to more complex industry structures.*

FIGURE 3

Under the standard quantity Cournot hypothesis on firm behavior, $dq_j/dq_i = 0$ for all $j \neq i$, and $\alpha_i = 0$. Therefore

THEOREM 2: *In a homogeneous product, quantity Cournot industry,*

$$\phi = \frac{1}{\epsilon} \sqrt{H} \qquad (10)$$

where $H \equiv \Sigma_{i=1}^{n} s_i^2$ is the Herfindahl Index of market concentration.[15]

This result can be interpreted in several ways. First, it establishes a precise welfare theoretic basis for and meaning of the popular H index. Second, it suggests that the Herfindahl Index should be modified, as in (10), to permit the interpretations afforded by Propositions 2, 4, and 5. However, we emphasize the caution in Proposition 3: while the H index may indicate the *sensitivity* of welfare performance to locally optimal industry adjustments, it does not indicate the *level* of welfare performance. Finally, (9) and Theorem 2 show that the welfare gradient significance of the Herfindahl Index rests crucially

[13]This follows from the convexity of $s_i^2(1 + \alpha_i)^2$ in s_i, which implies that $\Sigma s_i^2(1 + \alpha_i)^2$ is convex in the vector s. Then, ϕ is quasi convex in s since it is a monotone transformation of a function which is convex in s.

[14]Letting x denote the shift in share, consider $(1 + \alpha_i)^2(s_i + x)^2 + (1 + \alpha_j)^2(s_j - x)^2$ as a function of x. It is decreasing (increasing) in x iff a shift in share from j to i decreases (increases) ϕ. The derivative with respect to x, at $x = 0$, is $2(1 + \alpha_i)^2 s_i - 2(1 + \alpha_j)^2 s_j$. The statements in the text follow. Note, however, that a shift in share from a smaller firm to a larger firm can decrease ϕ if the smaller firm's α is enough larger in absolute value than the α of the bigger firm. Such paranoid delusions of grandeur on the part of small firms are usually ruled out by more explicit behavioral assumptions.

[15]Using a different approach, Trout Rader and others have shown that the Herfindahl Index follows from the Cournot assumption.

on the quantity Cournot behavioral hypothesis.

Other sets of assumptions on firm behavior yield other specializations of ϕ.[16] For example, suppose that only the largest m firms in the industry perceive $\alpha_i = 0$, while the remaining smaller firms act as price takers who equate $P = C_i'(q_i)$. In this scenario, the truncated Herfindahl Index emerges from (9).

THEOREM 3: *In a homogeneous product industry composed of a competitive fringe of small firms and m quantity Cournot larger firms,*

$$\phi = \frac{1}{\epsilon}\left(\sum_{i=1}^{m} s_i^2\right)^{1/2} \tag{11}$$

where $s_i > s_j$ for $i \leq m$ and $j > m$.

The standard m-firm concentration ratio can be derived from yet another scenario. Let the smallest $n - m$ firms be passive price takers, and suppose that the biggest m firms ignore the fringe and choose their output levels to maximize their joint profits. Then, for $i \leq m$

$$P - C_i' = (P/\epsilon)\sum_{i=1}^{m} s_i$$

and (9) yields

THEOREM 4: *In a homogeneous product industry composed of a competitive fringe of small firms and m jointly quantity Cournot profit-maximizing larger firms,*

$$\phi = \frac{\sqrt{m}}{\epsilon}\sum_{i=1}^{m} s_i \tag{12}$$

where $s_i > s_j$ for $i \leq m$ and $j > m$.

Theorems 2, 3, and 4 together enable a novel style of comparisons. Given ϵ and s as data describing the current industry structure, we can compare the values of the industry performance gradient index that would result from different visions of the behavioral modes of the industry's firms. Since

$$\left(\sum_{i=1}^{m} s_i^2\right)^{1/2} > \left(\sum_{i=1}^{k} s_i^2\right)^{1/2} \text{ for } m > k$$

we can conclude from (10) and (11) that ϕ is smaller as more firms are viewed as being on the competitive fringe rather than as Cournot quantity setters. Because

$$\sqrt{m}\sum_{i=1}^{m} s_i > \left(\sum_{i=1}^{m} s_i^2\right)^{1/2} \text{ for } m > 1$$

ϕ is larger if the nonfringe firms are viewed as joint profit maximizers than it is if they are assumed to each be independent Cournot quantity setters. For example, in an industry composed of a competitive fringe of small firms and four larger firms each having a market share of .2, (12) and (11) give values of ϕ of $1.6/\epsilon$ and $.16/\epsilon$, respectively.

In fact, the judgment about which behavioral mode prevails in an industry can have more influence on the calculated value of ϕ than has the distribution of market shares.[17] For example, suppose one industry has a wide competitive fringe and six equal-sized collusive firms with a total share of .6, while another industry is composed of four equal-sized quantity Cournot firms. Then the first industry has $\phi = 1.47/\epsilon$, while the second, even though it is far more concentrated, has a signficantly smaller $\phi = .5/\epsilon$.

Theorems 2–4 show that a choice among standard concentration indexes should not be made without conscious reference to underlying hypotheses on firm behavior. Different hypotheses may best describe the modes of conduct of firms in different industries. Then, comparison of values of ϕ may, for example, entail comparing one industry's modified four-firm concentration ratio with another's modified Herfindahl Index.[18]

[16] It is shown in Dansby's thesis that several other concentration indexes (for example, Entropy and Hall-Tideman) can be derived as special cases of the index ϕ.

[17] Of course, the distribution of shares may be important data for judgment about behavioral mode.

[18] This tends to cast doubt on the interpretation of studies by Duncan Bailey and Stanley Boyle, Marshall Hall and Nicolaus Tideman, and Robert Kilpatrick, which attempt to determine the "optimal" standard concentration-index by computing the values of each type of index for *all* industries, and comparing them with statistical techniques.

In view of Propositions 4 and 5, the assumed modes of conduct of an industry's firms can be critical in the cost-benefit analysis of intervention. Moreover, differences in industries' elasticities of demand can reverse the ranking based on concentration and conduct. The industry performance gradient index enables a quantification of these effects in support of the enlightened conventional wisdom that concentration per se should not be the only criterion for intervention.

IV. Intervention Costs, Metrics and New Indexes

The index ϕ is a weighted sum of the components of the welfare gradient vector. Our methodology calculates the weights as components of a unit vector pointing in the locally best direction of adjustment of industry outputs. Since the calculation of this direction depends on the metric employed, different metrics yield different sets of weights for the gradient components, and, hence, different index formulae. This observation raises the question of the economic interpretation of the metric.

Consider the roles played by the metric in the interpretation of ϕ afforded by Proposition 4. *Precise* use of this result would entail predicting the Δq induced by intervention and then calculating $\phi(q^0)\,\rho(\Delta q, P^0)$. The validity of this quantitative procedure does not depend on the metric having any economic significance. However, if Proposition 4 were used to prepare *rough* estimates of the upper bounds of the benefits of intervention, ϕ would be most useful if perceived "sizes" of different government actions were positively related to the metric evaluations of the corresponding output changes. This gives a loose economic criterion for the selection of a metric.

A related criterion is that the metric assign distances to various output changes that are monotonically increasing with the intervention costs required to effect the changes. The index ϕ measures the local potential welfare gain per unit of distance of quantity adjustment. If the criterion were satisfied, then comparisons of values of ϕ, as per Proposition 5, would yield local comparisons of gross potential welfare benefits per dollar of required intervention cost. If instead the criterion were satisfied within but not across industries, then each industry's value of ϕ could be meaningfully deflated by a measure of intervention cost per unit of distance of output adjustment. Either way, ϕ could be rigorously interpreted as a local indicator of the potential benefit-cost ratio of government intervention.

The base-price-weighted Euclidean metric underlying the index ϕ has two salient properties that bear on the question of whether it tracks intervention costs. First, this metric treats symmetrically equal sized adjustments (valued at base prices) in the outputs of firms of disparate sizes. If it is socially more costly to effect a large than a small *percentage* change in the output of a firm, then, for example, the Euclidean metric over percent output adjustments may be a better indicator of intervention costs:

$$\rho(\Delta q, q^0) = [\sum (\Delta q_i / q_i^0)^2]^{1/2} \tag{13}$$

This measure of distance is unitless and makes changes in the quantities of different commodities commensurable by comparing each of them to the corresponding base output. With derivations analogous to those noted in Sections I and II we establish that

THEOREM 5: *For the metric defined in (13), the industry performance gradient index at q^0 is*

$$\phi^s = [\sum (W_i(q^0) q_i^0)^2]^{1/2}$$

In a homogeneous product industry, where (8) obtains,

$$\phi^s = [P^0 Q^0/\epsilon]\,[\sum_{i=1}^{n} s_i^4 (1+\alpha_i)^2]^{1/2}$$

If the homogeneous product industry is:

(i) *composed of quantity Cournot firms, then*

$$\phi^s = [P^0 Q^0/\epsilon]\,[\sum_{i=1}^{m} s_i^4]^{1/2}$$

(ii) *composed of a competitive fringe of small firms and m quantity Cournot larger firms, then*

$$\phi^s = [P^0 Q^0/\epsilon]\,[\sum_{i=1}^{m} s_i^4]^{1/2}$$

where $s_i > s_j \ \forall \ i \leq m < j$

(iii) *composed of a competitive fringe of small firms and m jointly quantity Cournot profit-maximizing larger firms, then*

$$\phi^s = [P^0 Q^0/\epsilon] \left[\sum_{i=1}^{n} s_i\right]\left[\sum_{i=1}^{m} s_i^2\right]^{1/2}$$

where $s_i > s_j \ \forall \ i \leq m < j$

The new index ϕ^s is in money units and has the properties cited in Propositions 1–5, with $\rho(\Delta q, P^0)$ being replaced by $\rho(\Delta q, q^0)$. There are two important distinctions between ϕ and ϕ^s. First, due to the units of the underlying metrics, ϕ^s increases with the value of industry output while ϕ is independent of this value. Second, holding other data fixed, ϕ^s is more sensitive than is ϕ to market share inequality because its underlying metric is itself sensitive to the size distribution of firms. This is a desirable feature of ϕ^s if the social costs of inducing a given absolute quantity adjustment are smaller for a large firm than they are for a firm with a small market share.[19]

Both Euclidean metrics have the property that for $\rho(\Delta q, \cdot) > 0$, $\partial\rho/\partial|\Delta q_i|$ is positive if $|\Delta q_i| > 0$ and is zero if $\Delta q_i = 0$. Consequently, the locally best direction of output changes has $|\Delta q_i| > 0$, unless $W_i(q^0) = 0$. Such a metric can be a monotonic transformation of the social costs of intervention only if these costs have the property near q^0 of decreasing returns to the scale of $|\Delta q_i|$.[20] Here, decreasing returns means that the additional social cost of intervention needed to induce the change $|\Delta q_i|$ in the output of firm i, averaged over $|\Delta q_i|$, is increasing with the size of the change.[21] Given sufficiently increasing average incremental intervention costs, an efficient allocation of a fixed budget for intervention activities in a particular industry will include actions to affect the outputs of every firm i with $W_i \neq 0$. If, on the other hand, there were sufficiently increasing returns, then a given budget would be best allocated to actions (presuming their existence) which change the output of only one firm. With constant average incremental intervention costs, the optimal set of firms forced to adjust would depend on the size of the allowed budget, but a small budget would be best focused on a single firm.

The city-block metric[22] with base-price weights $\rho^*(\Delta q, P^0) = \Sigma_i P_i^0 |\Delta q_i|$, reflects constant intervention costs in that $\partial\rho^*/\partial|\Delta q_i| = P_i^0$ irrespective of Δq. With this metric, the locally best direction change adjusts only that q_k with the largest value of $|W_k(q^0)|P_k^0$.

THEOREM 6: *For $\rho^*(\Delta q, P^0)$, the industry performance gradient index at q^0 is*

$$\phi^* = \max_i |W_i(q^0)/P_i^0| = \max_i |\{P_i^0 - C_i'(q_i^0)\}/P_i^0|$$

In a homogeneous product industry:

(i) *composed of quantity Cournot firms and possibly a competitive fringe of smaller firms, $\phi^* = (1/\epsilon)\max_i s_i$.*

(ii) *with the largest m firms joint quantity Cournot profit maximizers and with*

$$1 + \alpha_j < \frac{1}{s_j}\sum_{i=1}^{m} s_i$$

[19]There is another interesting distinction between ϕ and ϕ^s in monopolistically competitive industries. If each firm's demand depends on the prices of its rivals and if each firm is a price Cournot profit maximizer, then, it is shown in the authors that $\phi = [\Sigma(1/\epsilon_i)^2]^{1/2}$ while $\phi^s = [\Sigma P_j^0 q_j^0] \times [\Sigma(r_i/\epsilon_i)^2]^{1/2}$. Here, ϵ_i denotes the own-price elasticity of demand for the ith firm's product, and $r_i = P_i^0 q_i^0/\Sigma P_j^0 q_j^0$ is the revenue share of the ith firm. In this context, while ϕ does not have the appearance of a concentration index, ϕ^s has the elements of a Herfindahl Index, with revenue shares replacing the usual quantity shares.

[20]This would be the case, for example, if inexpensive "jawboning" could achieve small output adjustments, while firms were increasingly intransigent about yielding to government requests for larger output changes.

[21]Viewing the cost of intervention as the cost of "producing" the vector Δq, this property is not the standard notion of multiproduct decreasing returns to scale. Instead, it pertains to the behavior of average *incremental* costs; i.e., to the returns to the scale of Δq_i, holding the other components of the Δq vector fixed. See John Panzar and Willig (1977a, p. 11; 1977b).

[22]This metric measures the distance between two points as the sum of the absolute differences between their coordinates. For example, in a city whose streets form a grid, the metric gives the physical distance that is travelled in going from one point to another along streets.

for each smaller firm j,

$$\phi^* = (1/\epsilon) \sum_{i=1}^{m} s_i$$

The new index ϕ^* is unitless and has the properties cited in Propositions 1–6. Part (ii) of Theorem 6 shows that ϕ^* reduces to the m-firm concentration ratio, given that the smaller firms do not have exceedingly large values of α_j. In particular, it suffices with $m \geq 2$ that $\alpha_j < 1$ for each noncolluding firm. This condition is considerably weaker than the assumption of price-taking ($\alpha_j = -1$) smaller firms, which is needed to reduce ϕ to the m-firm concentration ratio. Here, the smaller firms are ignored in ϕ^* because the optimization under the constant returns metric adjusts only the output of the worst offending firm. It is this same feature of ϕ^* which makes it depend on only the largest market share in a quantity Cournot industry. In contrast, the Herfindahl Index emerges from ϕ in the same behavioral scenario because the increasing average incremental intervention costs embodied in the Euclidean metric makes it optimal to adjust the outputs of all noncompetitive firms.

The ranking of industries given by the performance gradient index can depend on the metric employed to derive the index. For example, suppose industry A has the vector of market shares $\{.35, .20, .15, .30\}$, while industry B is composed of three equal-sized firms. If both A and B are quantity Cournot industries with the same observed price elasticity of demand, then ${}^A\phi^* = .35/\epsilon > .33/\epsilon = {}^B\phi^*$. However, ${}^A\phi = .52/\epsilon < .57/\epsilon = {}^B\phi$. Moreover, since ${}^A\phi^S$ is $.16/\epsilon$ times the value of the output of industry A and ${}^B\phi^S$ is $.19/\epsilon$ times the value of the output of industry B, the ranking derived from the Euclidean metric over percentage output changes depends largely on the relative values of the outputs of the industries.

V. Concluding Remarks

We have presented an analytic framework for deriving industry performance gradient indexes as indicators of the rate of potential improvement in industrial welfare performance. The indexes combine data on the current structure of the industry with measures of the conduct of the industry's firms. Thus, they allow a quantitative synthesis of elements that bear on the welfare gains from government intervention.

The manner in which these elements combine depends upon characteristics of the social costs of intervention. In particular, the returns to scale properties of intervention costs and the dependence of these costs on firm size emerge as critical factors that warrant additional investigation.

Our analysis provides welfare-theoretic interpretations of the standard concentration indexes and enables the construction of new ones from explicit assumptions. Most importantly, the framework forces the placement of the analytic focus on firm behavior and the social costs of intervention.

REFERENCES

M. A. Adelman, "Monopoly and Concentration: Comparisons in Time and Space," in Tullio Bagiotti, ed., *Essays in Honor of Marco Fanno,* Cedam, Italy 1966.

D. Bailey and S. E. Boyle, "The Optimal Measure of Concentration," *J. Amer. Statist. Assn.,* Dec. 1971, *66,* 702–06.

John M. Blair, *Economic Concentration,* New York 1972.

K. Cowling and M. Waterson, "Price-Cost Margins and Market Structure," *Economica,* May 1976, *43,* 267–74.

R. E. Dansby, "Welfare Economic Implications of Some Pricing, Capacity Investment and Advertising Decisions," unpublished doctoral dissertation, New York Univ. 1976, ch. 4.

———, **and R. D. Willig,** "Industry Performance Gradient Indexes," econ. disc. paper, Bell Laboratories 1976.

J. B. Delaney, "On The Measurement of Concentration: A Cross-Sectional Analysis of Alternative Indices," unpublished doctoral dissertation, Boston College 1974.

M. Hall and N. Tideman, "Measures of Concentration," *J. Amer. Statist. Assn.,* Mar. 1967, *62,* 162–68.

O. C. Herfindahl, "Concentration in the Steel Industry," unpublished doctoral disserta-

tion, Columbia Univ. 1950.

A. C. Johnson and P. Helmberger, "Price Elasticity of Demand as an Element of Market Structure," *Amer. Econ. Rev.*, Dec. 1967, *57*, 1218–21.

R. W. Kilpatrick, "The Choice Among Alternative Measures of Industrial Concentration," *Rev. Econ. Statist.*, May 1967, *49*, 258–60.

A. P. Lerner, "The Concept of Monopoly and the Measurement of Monopoly Power," *Rev. Econ. Stud.*, June 1934, *1*, 157–75.

J. E. Morton, "A Student's Guide to American Federal Government Statistics," *J. Econ. Lit.*, June 1972, *10*, 371–97.

J. C. Panzar and R. D. Willig, (1977a) "Free Entry and the Sustainability of Natural Monopoly," *Bell J. Econ.*, Spring 1977, *8*, 1–22.

______ **and** ______, (1977b) "Economics of Scope, Product Specific Economies of Scale, and Competitive Equilibrium," unpublished paper, Bell Laboratories 1977.

R. C. Parker, "Comparable Concentration Ratios for 213 Manufacturing Industries Classified by Producer and Consumer Goods and Degree of Product Differentiation," Bureau Econ., Federal Trade Commission, Washington 1967.

Trout Rader, *Theory of Microeconomics*, New York 1972.

K. W. Rothschild, "The Degree of Monopoly," *Economica*, Feb. 1942, *9*, 24–39.

Paul A. Samuelson, *Foundations of Economic Analysis*, Cambridge, Mass. 1947.

F. M. Scherer, *Industrial Market Structure and Economic Performance*, Chicago 1970.

T. Scitovsky, "Economic Theory and the Measurement of Concentration," in *Business Concentration and Price Policy*, Universities-Nat. Bur. Econ. Res. conference series, Princeton 1955, 101–13.

E. Silberberg, "A Revision of Comparative Statics Methodology in Economics," *J. Econ. Theory*, Feb. 1974, *7*, 159–72.

Eugene M. Singer, *Antitrust Economics*, Englewood Cliffs 1968.

The "Stationarity" of Shadow Prices of Factors in Project Evaluation, with and without Distortions

By JAGDISH N. BHAGWATI AND HENRY WAN, JR.*

Until recently, the literature on cost-benefit analysis for projects has been largely within the domain of research on "public monopoly," literature currently reviewed by Jacques Lesourne, (ch. 3), and the work of public finance theorists as typified in the celebrated practical work of Ian Little and James Mirrlees in their *Manual*, and in the recent theoretical contribution of Peter Diamond and Mirrlees. International trade theorists have, however, turned now to the analysis of these problems, starting with the early work of Vijay Joshi and Deepak Lal, then that of W. M. Corden, and most recently culminating in the contributions of Ronald Findlay and Stanislaw Wellisz, and T. N. Srinivasan and Bhagwati.

The work of Findlay-Wellisz and Srinivasan-Bhagwati (F-W-S-B) explicitly deploys the tools, insights and ideas of general equilibrium international trade theory. In particular, their analyses have been addressed to the question of deriving the shadow prices for primary factors for the purpose of project evaluation in the presence of distortions: Findlay-Wellisz (F-W) considering product-market and trade distortions and Srinivasan-Bhagwati (S-B) also extending their analysis to a number of factor-market distortions.

Their analyses has been conducted essentially within the framework of the two-by-two small-country model of traditional international trade theory. An important consequence is what might be called the "stationariness" of the "marginal variational" shadow prices of factors (derived by marginal, i.e., infinitesimal, variation) such that, as S-B phrased it,

> [W]hile we have confined our analysis to 'small' projects, drawing infinitesimal resources away from the existing distorted situation, it is equally clear from our analysis that the results will also hold for 'large' projects. Given the Rybczynski-line properties of the different models, the shadow prices of factors will be identical for small and large shifts of factors into the project. [p. 113]

When the Rybczynski-line properties no longer hold, the marginal variational shadow prices applicable for single projects with infinitesimal factor withdrawals will indeed vary as the factor endowment vector varies. Similarly, for a project withdrawing finite amounts of factors[1] the marginal variational shadow prices computed *before* the withdrawal will then differ from those computed *after* the withdrawal. Moreover, shadow prices computed by marginal variations from the "residual factor vector" (after the withdrawal) will depend upon the size and composition of the factors withdrawn.[2]

The "stationarity" of the marginal varia-

*Professor of economics, Massachusetts Institute of Technology, and professor of economics, Cornell University, respectively. Our thanks are due to Trent Bertrand, Simone Clemhout, Peter Diamond, Ronald Findlay, Earl Grinols, T. N. Srinivasan, Lance Taylor, the managing editor of this *Review*, and an anonymous referee for helpful suggestions. Comments at seminars at Columbia, Yale, Johns Hopkins, and the World Bank also helped greatly. Partial support under National Science Foundation Grant SOC77-07188 is gratefully acknowledged.

[1]Equally, for a successive sequence of "small" projects, collectively withdrawing finite amounts of factors in the aggregate.

[2]As a matter of pure formality, "true" shadow prices can be defined, a posteriori, for projects withdrawing finite factor dosages (see the authors, Appendix I). Such shadow prices, used for project evaluation, *will* tautologically yield the opportunity cost. However, these prices will vary from project to project and their derivation will require each time the solution of a full programming problem for project selection. Such shadow price computation will therefore become a purely academic exercise: the projects having been selected in the programming problem already. A similar point has been made by Bhagwati and Srinivasan in relation to estimating the

tional shadow prices, in the presence of *non-infinitesimal* factor withdrawals, such that the valuation of these factors in project use at the marginal variational shadow prices nonetheless equals their true social opportunity cost, is therefore a critical question.[3] This is precisely the issue, effectively skirted in the standard analyses of shadow prices in project evaluation by the convenient assumption of "small" projects, that we propose to examine in the present paper. Towards this end, we propose to relax the two-by-two property of the F-W-S-B model to allow for many goods and factors: it is shown that uniqueness and stationarity of the marginal variational shadow prices are not always guaranteed once the number of goods differs from that of factors.

Section I recapitulates the basic F-W-S-B analysis, retaining the two-by-two model but distinguishing between the with-distortion and the no-distortion cases. Section II examines the many-goods-and-factors cases:

Case 1: *Goods equal factors, no distortion.*

Case 2: *Goods equal factors, with distortion.*

Case 3: *Goods outnumber factors, no distortion.*

Case 4: *Goods outnumber factors, with distortion.*

Case 5: *Factors outnumber goods, no distortion.*

Case 6: *Factors outnumber goods, with distortion.*

Section III offers concluding observations, indicating the applicability of our analysis to other problems in trade theory (for example, the transfer problem and the welfare effects of labor mobility) and the relationship of our results to mathematical programming. Owing to lack of space, we do not report here: 1) how the replacement of the ad valorem tariff distortion by a *quantitative* quota distortion would destroy the stationarity of shadow prices in the F-W-S-B model; nor 2) how the major propositions in Section II can be proved with rigor, and generalized to cover the cases of (i) joint outputs, (ii) traded inputs, (iii) nontraded, domestically produced inputs unfit for consumption, (iv) primary inputs with variable supply, and (v) consumable nontraded domestic products. Interested readers may refer to the authors (Section II; Appendices I, II).

I. Recapitulating and Completing Analysis within the F-W-S-B Model

The F-W-S-B model is characterized by three key features: constant-returns-to-scale production functions;[4] two primary factors producing two traded goods; and fixed foreign prices for the two traded goods (the Samuelson "small-country" assumption). The problem of deriving shadow factor prices for a project producing a third traded good then is tantamount to deriving the changes in outputs of the two traded goods (x_1 and x_2) that follow from the withdrawal of factors (v_1 and v_2) from existing allocations and then evaluating these output changes at (the fixed) *international* prices. According to the Little-Mirrlees "rule," the shadow price of a factor is precisely the value of output foregone when this factor is marginally, that is, infinitesi-

effective rate of protection (*ERP*) index for resource allocation prediction in the case of generalized factor substitution. They argue that, in general, to compute the "correct" *ERP* index, we must solve the general equilibrium system for the tariff change; but if we have done that, we already know the total resource allocation change and we do not need the *ERP* index to tell us the direction of such change.

[3]For an extremely scathing and articulate critique of cost-benefit analysis by a programmer-planner who argues that shadow prices which apply to negligible (i.e., infinitesimal) projects are of negligible interest, see Asok Rudra. He is clearly assuming what we christen here the "nonstationarity" of the marginal variational shadow factor prices and therefore his critique, while fundamentally sound in principle, goes too far in failing to show awareness of possible stationarity of shadow prices.

[4]F-W and S-B do not explicitly rule out factor-intensity reversals. On what happens when they are present, see our analysis below.

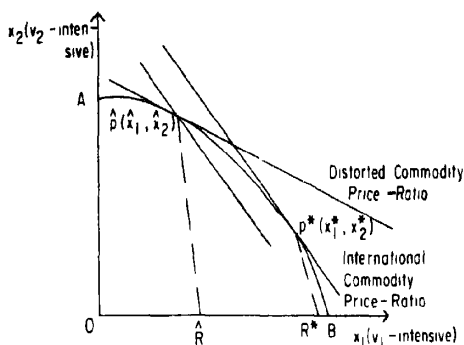



FIGURE 1

mally, withdrawn at the distorted market prices, with the valuation itself being carried out at the international prices.[5]

Thus, take Figure 1 which depicts the usual production possibility curve for the two traded goods, x_1 and x_2. Distinguish now two cases: 1) no distortion; and 2) a (given) product-market or trade distortion.[6] For the former case, production will take place at P^*; for the latter case, it will occur at $\hat{P}$. In the former case, the withdrawal of factor v_2, assumed to be intensively used in commodity x_2, will successively reduce the output of x_2 and increase it for x_1 along the Rybczynski line P^*R^* whereas a similar withdrawal of factor v_2 will define, in the latter case, a similar (nonparallel) Rybczynski line $\hat{P}\hat{R}$.[7] In each case, the commodity price ratio is clearly held constant as the factor is withdrawn for use in the project producing commodity x_3; in the no-distortion case this is the international price ratio whereas, in the distortion case, it is the distorted commodity price ratio. The *evaluation* of the changes in outputs along the relevant Rybczynski line, however, is at the *international* prices in *both* cases.[8]

An important consequence is that, as long as the withdrawal of factors, no matter how large, permits the small economy to remain on the Rybczynski line in Figure 1, whether there is a given distortion or none, the shadow factor prices will be "stationary."[9] What happens if factors are withdrawn such that the economy moves off the Rybczynski line?

First, focusing on the *no-distortion* case, assume that the withdrawal of primary factor v_2 leads the economy finally to complete specialization on commodity x_1 at the bottom of P^*R^* at R^*. If then another unit of v_2 is withdrawn, it is evident that the economy will move towards the origin along the horizontal axis and that, as it does so, the real rental of factor v_2 will rise in terms of *both* of the traded goods, x_1 and x_2. Essentially, one has slipped out of the "linearity" property of the system and diminishing returns to varying proportions are now taking over. The net result therefore is that, for any withdrawals of the primary factor v_2 beyond those that lead from P^* to R^*, the shadow prices of the

[5]Given the small-country assumption, this rule leads to the correct choice of a project because if the social cost of factors used therein, so derived, is below the international valuation of the project's output, the acceptance of the project will increase the international valuation of total output in the economy. When that happens, as long as the degree of consumption distortion remains unchanged (as when a tariff is in place), the analyst can infer that social welfare (as conventionally defined) will have increased as well (except in the presence of inferior goods when multiple equilibria can arise and society may be at an inferior equilibrium although a superior equilibrium is always available if the international valuation of total output has increased). For a full discussion of these issues, and the relationship between the Little-Mirrlees rule and the Diamond-Mirrlees technique for deriving identical shadow prices, see Srinivasan and Bhagwati.

[6]A product-market distortion essentially implies that the domestic producers face production tax-cum-subsidy-inclusive prices or tariff-inclusive prices, rather than the international prices for the two goods. If we invoke a tariff, there will be a consumption distortion as well. But this creates no difficulties for our analysis, so that our analysis holds equally for trade distortions.

[7]The slope of the Rybczynski lines reflects the average productivity of the primary factors whereas the slope of the commodity price ratio at which the Rybczynski line is defined reflects, of course, the marginal productivity of the factors. On this, as also on the relationship between the Rybczynski lines at different points on the production-possibility curve (as at P^* and $\hat{P}$), see Trent Bertrand and Frank Flatters, and Richard Brecher in particular.

[8]For simplicity, Figure 1 has been drawn such that the possibility of a negative shadow factor price is ruled out. On this paradoxical possibility, see Srinivasan and Bhagwati, and Bhagwati, Srinivasan, and Wan.

[9]S-B do not consider what happens beyond the Rybczynski line. Rather they consider other questions such as the relationship of their results on shadow factor prices to the use of domestic resource costs and effective rate of protection as project evaluation criteria.

factors are no longer stationary and can be shown to increase (for the factor withdrawn) for successive withdrawals of the factor for utilization in the project in question. This also implies that, for these ranges of increasing cost withdrawal of factor v_2, the use of the shadow prices for marginal variation at P^* would yield an *understatement* of the true shadow cost of the factor. Alternatively, one may phrase this to say that the use of marginal variational shadow prices, when stationariness does not obtain, ignores the "secondary cost" that must be added to the "primary cost" as measured at the marginal variational shadow prices.

All this can be seen perfectly generally, for withdrawals of *both* factors, in terms of the McKenzie-Chipman diversification cone in Figure 2. Assuming that q_1 and q_2 define the v_1/v_2 ratios chosen at the prevailing commodity price ratio (i.e., at P^* in Figure 1) and the associated factor-price ratio which is the slope of the line q_1q_2, note that the overall factor endowment ratio must be a weighted sum of the two sectoral factor proportions.[10] It is evident then that q_1Oq_2 defines a diversification cone: as long as the factor-endowment vector lies strictly *within* the cone, both goods will be produced at the postulated commodity and associated factor-price ratios. Hence the aggregate endowment vector $(\bar{v}_1, \bar{v}_2)$ is shown to lie within the cone q_1Oq_2, indicating production of both goods at levels x_1^* and x_2^* (corresponding to P^* in Figure 1). Note that the aggregate endowment vector $(\bar{v}_1, \bar{v}_2)$ is shown by the *parallelogram* in Figure 2 to be a weighted vector sum of the factor proportions in the two goods, the actual outputs being read off from the isoquants x_1^* and x_2^*.

It follows immediately that (given the commodity price ratio) as long as the *residual* factor vector, left over after successive withdrawals of the factors for project use, continues to lie within the cone q_1Oq_2, the equilibrium factor-price ratio need not change from

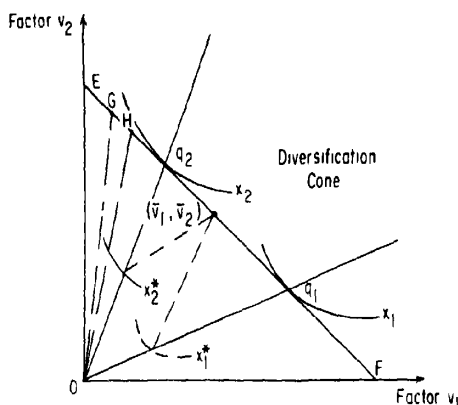



FIGURE 2

q_1q_2 and all changes in factor endowments will be accommodated merely by changes in the composition of outputs. Hence, the McKenzie-Chipman diversification cone *in the present case* is identical to what we shall refer to as the *Rybczynski cone*: the latter being the set of residual factor vectors for which marginal variational shadow factor prices will remain stationary at the slope of the price line q_1q_2. This stationarity of these shadow factor prices will disappear, however, as soon as the residual factor vector slips out of this cone.[11] Note also the following:

(i) As must be evident, the slope of the line q_1q_2 represents the ratio of the marginal variational shadow factor prices, these being the prices corresponding to infinitesimal factor withdrawal.

(ii) The secondary cost of withdrawn factors that leave the residual factor vector outside the Rybczynski cone may be defined as the excess of their true shadow cost over their cost if measured at the marginal variational prices. The *proportionate* secondary cost, defined as the ratio of the secondary cost to the cost at the marginal variational prices, will be an increasing function of the extent to

[10]Thus, denoting the overall endowments as $\bar{v}_1$ and $\bar{v}_2$, we must have (assuming full employment) $\bar{v}_1/\bar{v}_2$ as the v_2 share-weighted sum of the v_1/v_2 ratios in the production of the two goods.

[11]Reverting to Figure 1, note then that all withdrawals of factor v_1 resulting in moves along P^*R^* leave the residual factor vector in the diversification cone whereas further withdrawals of v_1, leading to moves along R^*O, imply that the residual factor vector has slipped out of the cone.

which the residual factor vector slips out of the Rybczynski cone.[12]

(iii) Finally, note that the unique Rybczynski cone in Figure 2 implicitly rules out factor-intensity reversals. If we had allowed for such reversals, however, we would have had more than one diversification cone. In this case, even if production were no longer specialized, secondary cost would emerge. Moreover, as before, the more the residual factor vector slips out of the Rybczynski cone, the higher will be the share of secondary cost in the entire project cost.

Second, for the *with-distortion* case, the preceding analysis for the no-distortion case holds qualitatively with one exception: the secondary cost can be a secondary gain. Returning to Figure 2, note that in the distortion-free case, for residual factor vectors that take the economy outside the Rybczynski cone $q_2 0 q_1$, the true social cost of factors in project use will be understated because valuing, say, a residual endowment at H at the marginal variational prices overstates its true value and hence understates the difference between the preproject and the postproject valuation of the quantities of commodities x_1 and x_2 that are being produced: this being the secondary cost noted above. But suppose now there is an export tax on x_2 so that its price is lower on the domestic market than on the world market. Guided by the domestic prices, firms collectively will produce more x_1 and less x_2 than what it takes to maximize the national output at international prices. Now assume a withdrawal of inputs which causes the residual factor vector to leave the diversification-cum-Rybczynski cone, so that good x_2 alone is produced. This change of the output mix improves on the original distortion-caused misallocation and hence increases the internationally valued national output. Conceivably, this effect may outweigh the secondary cost from the law of variable proportions and result in a net secondary gain.[13]

[12]Thus, for residual factor endowments lying on the ray $0G$ in Figure 2, the proportionate secondary cost will be higher than for those lying on the ray $0H$.

II. Many Goods and Factors, with and without Distortions

The uniqueness and stationarity of the marginal variational shadow factor prices in the two-by-two F-W-S-B model, with and without the specified distortions, do not necessarily carry over to the cases with unequal numbers of goods and factors that need to be analyzed as soon as we consider many goods and factors. The analysis of these cases is simply developed in this section, at a qualitatively insightful level.[14]

The precise questions that are addressed for the six possible cases discussed here are noted best by recalling that the shadow price of factors withdrawn for project use is the sum of the resulting changes in outputs valued at *international* prices. Write this shadow cost as

$$c_k = -p^* \Delta_k x$$

where c_k denotes the shadow factor cost of a project (program) k, p^* is the vector of international prices, and $\Delta_k x$ is the vector of changes in nonproject production in the country as a result of the withdrawal of factors for project use. (Here x, p, and p^* represent m-dimensional vectors denoting output quantities, (domestic) output prices facing the firms, and international prices, respectively. Also v, w, and w^* will be n-dimensional vectors representing factor quantities and factor prices related to domestic and international prices, respectively.) Noting that

$$c_k (= -p^* \Delta_k x) = -w^* \Delta_k v$$

for preproject marginal variational shadow factor prices w^* and infinitesimal factor withdrawal for project use, we can then pose three questions of principal importance in this

[13]The secondary gain in fact may turn the social cost of the factors withdrawn for project use into a net gain, resurrecting the paradox of negative shadow prices of factors noted in Bhagwati, Srinivasan, and Wan.

[14]The algebraic analysis, necessary for general formulations, is in Appendix 1 of our working paper; available from the authors on request.

section:

(i) Can the social opportunity cost c_k be *uniquely* defined?

(ii) Will the shadow factor prices be stationary for finite factor withdrawals in the perfectly general sense that $-w^*\Delta_k v$ will equal $(c_k =) - p^*\Delta_k x$, i.e., the valuation of the withdrawn factors at the preproject marginal variational shadow prices will equal the social opportunity cost of the foregone output?

(iii) Will the use of marginal variational shadow factor prices, when inappropriate, necessarily understate the social cost of the project?[15]

The results of our analysis are summarized in advance for the reader's convenience in Table 1.

TABLE 1—SIX ALTERNATIVE CASES: OUTCOMES REGARDING MARGINAL VARIATIONAL SHADOW FACTOR PRICES[a]

Cases	No Distortion	With Distortion
Goods Equal Factors	I. Stationarity	II. Stationarity
Goods Outnumber Factors	III. Stationarity	IV. Shadow Prices May be Undefined
Factors Outnumber Goods	V. Possible Nonstationarity	VI. Possible Nonstationarity

[a]"Extreme" cases noted in fn. 15 are excluded. For the no-distortion cases, the use of marginal variational shadow factor prices will necessarily underestimate the true social cost of a project; not so for the with-distortion cases.

A. *Cases 1 and 2: Equal Numbers of Goods and Factors, with and without Distortions*

The introduction of more than two goods and factors, as long as goods are equal in number to factors, changes the results of the two-by-two F-W-S-B model in no essential manner. As before, a diversification cone can be defined and, since it is also the Rybczynski cone, the marginal variational shadow factor prices are unique and also stationary as long as the residual factor endowment leaves the economy within this cone.

When the residual factor endowment takes the economy into the nonstationarity zone, the use of marginal variational shadow factor prices will necessarily understate the true social cost of the factors used in the project if distortions are absent; but it may overstate the true cost (i.e., the secondary cost may turn into a secondary gain) when distortions are present.

B. *Case 3: Goods Outnumber Factors, No Distortions*

It is equally evident that, where goods outnumber factors and there is no distortion, there will be no problem with uniqueness and stationarity in general. A diversification-cum-Rybczynski cone can again be defined and, within it, the marginal variational shadow factor prices will be unique and stationary. The product mix will be indeterminate of course, but this does not affect the shadow factor prices, as should be evident by redrawing the Lerner-Findlay-Grubert diagram (Figure 2) for more than two goods by merely putting in more isoquants tangent to the linear segment q_2q_1.

The use of marginal variational shadow factor prices for evaluating the social cost of

[15]In answering these questions, we will generally ignore the "extreme" cases which provide exceptions to our propositions and analysis, so that we can avoid frequent repetition of caveats regarding them. Such extreme cases are reviewed readily for the two-by-two model Thus (a) certain projects will cause difficulties. For instance, consider a project using up resources of magnitude and proportions such that the residual vector is no longer in the same diversification cone as the initial vector. Either complete specialization will occur, or the residual vector may now be in another diversification cone (as may happen when factor intensity is reversible), and shadow prices will change (b) Certain projects almost surely will cause no difficulty. Thus, consider the case where the diverted resources, and hence the residual resources, are proportional to the initial resources. (c) Certain circumstances almost surely will cause difficulties. For instance, take the case where the two unit-output isoquants, one for each project, are tangential to a common tangent at the same point. (d) Certain circumstances almost surely will cause no difficulties. For instance, consider the case where, over some relevant range, the two inputs are "perfect substitutes" in the sense that a linear segment of the unit-output isoquant of at least one product prevails.

the factors used in the project, when the residual factor endowment leaves the economy outside this cone, will necessarily understate their true social cost: that is, secondary cost will necessarily arise.

C. *Case 4: Goods Outnumber Factors, with Distortions*

While, however, there is no problem as long as goods outnumber factors in the absence of distortions, the introduction of distortions leads to difficulties. Intuitively, it is easy to see why. For the diversification cone is defined with respect to p, the *domestic* output prices. Hence, within the cone, the indeterminacy of the output mix is still compatible with unique opportunity cost and hence with unique and stationary factor valuations as long as the marginal variational changes in outputs from factor withdrawals are evaluated at the *domestic* prices. However, the shadow factor prices require these output changes to be evaluated at the *international* goods prices p^*. When $p = p^*$ (i.e., Case 3), shadow factor prices will also be unique and stationary within the diversification cone.

But when $p \neq p^*$, the social opportunity cost of the withdrawn factors, even when the residual vector remains in the diversification cone, may not be unique but will reflect the particular product mix happening to obtain out of the indeterminate many. Hence the associated shadow factor prices may be undefined.

Suppose, however, that a "planner" *chooses* for each residual vector that particular output mix whose value, evaluated at international prices p^*, is *maximal*. Could we then argue that in this event the shadow factor prices will be stationary within the diversification cone? The answer unfortunately is again in the negative, generally speaking. Note that even if it were in the affirmative, to use such shadow prices for project evaluation, we would have to assume that these maximal-value product mixes were *in fact* the equilibrium output mixes obtaining in the economy before and after the factor withdrawals for the projects in question. Otherwise, the "true" opportunity costs of the withdrawn factors would not correspond to the "shadow" opportunity costs as calculated with the maximal-value procedure.

All this should be perfectly intuitive, once the difference between p and p^* under the specified distortion is grasped. It can be established more formally[16] but may rather be illustrated to great advantage with the aid of the "micro-theoretic" diagrams, Figures 3A–E. Figures 3A–C introduce the basic technique, while Figures 3D–E illustrate our basic propositions with this technique by using a three-good, two-factor, with-distortion depiction.

In Figure 3A, two inputs v_1 and v_2 are used to produce output x, which is depicted along the vertical axis $0x$. At the initial endowment v, height $v\bar{v}$ of the production surface $0CD$ shows the output value before factor withdrawal for project use. A project using both factors in proportion to the initial endowments will leave a residual vector $0B$ within the ray $0\bar{v}$. The difference in height between $B\bar{B}$ and $v\bar{v}$ reflects then the project cost. The unique tangent plane $0PP'$ to surface $0CD$ at $\bar{v}$ represents clearly the marginal variational valuation of any residual factor vector.[17] Since $\bar{B}$ lies both on $0C\bar{v}D$ and on its tangent plane $0PP'$, such valuation incurs no secondary cost for the project in question. By contrast, a project using factors *not* proportional to the initial endowments will leave a residual vector $0A$ off the $0v$ ray (which is of course the Rybczynski cone in the present case). The maximum output producible at A is the height of the $0CD$ surface $A\tilde{A}$, which is less than the marginal variational valuation $A\bar{A}$ for $0A$. The overvaluation of the residual vector thus causes an understatement of the true opportunity cost of (the factors used in) the project.

Figure 3B extends this construction to incorporate two goods x_1 and x_2. If we select units such that the output prices are unity for both goods, the surface $0C\bar{q}_1$ and $0D\bar{q}_2$ reflect what is producible if all inputs are used to

[16]Mathematical proofs are contained in the unpublished Appendix I in our working paper.

[17]The directional cosines at v reflect the marginal products of the two inputs. Their ratio is the slope of line CD in plane $v_1 0 v_2$, while $PP' // CD$.

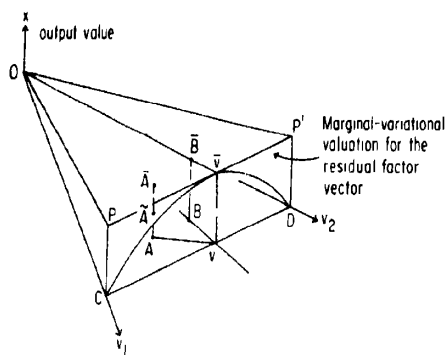



FIGURE 3A

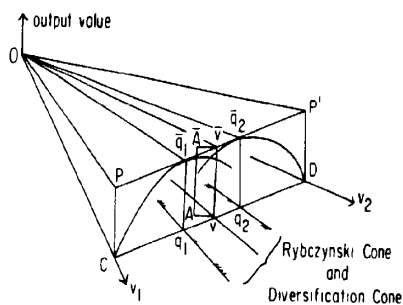



FIGURE 3B

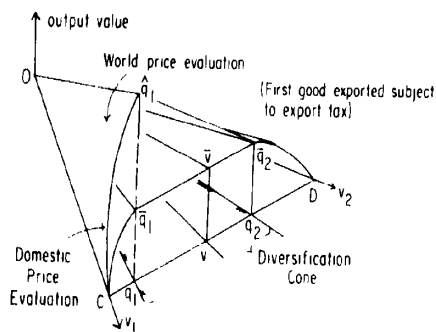



FIGURE 3C

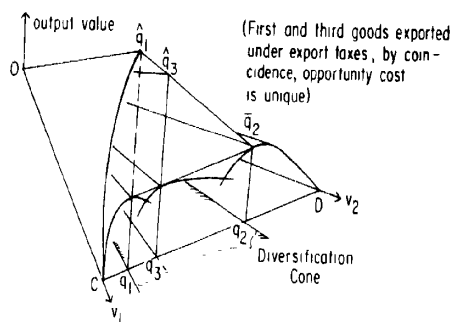



FIGURE 3D

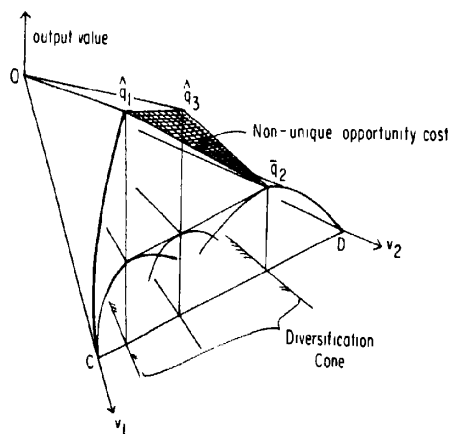



FIGURE 3E

produce only x_1 and x_2, respectively, and their convex hull therefore has surface OCD and tangent plane OPP' reflecting the marginal variational valuation of all residual factor vectors. The factor proportions producing good x_1 (x_2) under the prevailing factor prices before the project are reflected by ray Oq_1 (Oq_2). If a project leaves a residual factor vector OA within the Rybczynski cone q_1Oq_2, then the marginal variational valuation $A\underline{A}$, thereof, agrees with the true maximum output value producible therewith: there is therefore no overvaluation for OA and thus no understatement of the project cost.

Figure 3C modifies Figure 3B to allow international prices to differ from domestic prices, illustrating the distortion. Thus, it portrays an export tax on commodity x_1. Consequently, the world market value of any output mix is higher than the domestic market value if and only if some x_1 is produced. The $OC\hat{q}_1\bar{q}_2D$ surface is constructed to reflect this fact. The $O\hat{q}_1\bar{q}_2$ planar segment reflects therefore the true social value of any residual factor vector falling within the diversification cone q_1Oq_2.

We can now proceed to our full illustration of the three-good, two-factor, with-distortion case. Thus Figures 3D–E additionally portray output of x_3 which is also assumed (like x_1) to be exported under an export tax. It is further assumed that the factor intensity of good x_3 lies between that of the other two goods. Clearly the construction of the $OC\hat{q}_1\hat{q}_3\bar{q}_2D$ surface is similar to that of the $OC\hat{q}_1\bar{q}_2D$ surface in Figure 3C.

Figure 3D then shows the case where, by coincidence, $\hat{q}_1\hat{q}_3\bar{q}_2$ fall on one line such that $O\hat{q}_1\hat{q}_3\bar{q}_2$ constitutes a single planar segment whose height at any residual factor vector within the diversification cone Oq_1q_2 then reflects the true social (i.e., international) value of that vector. Hence also the difference between that height and the height $v\bar{v}$ (corresponding to the preproject endowment vector) represents the true opportunity cost of the factors withdrawn for project use. Clearly, because of the planar segment, this opportunity cost is unique even within the diversification cone.

By contrast, the general case is depicted in Figure 3E. Here the planar segment disappears and instead one gets indeterminate, social opportunity cost for the project within the diversification cone.[18] Depending on the output mix, we can see that the (social) value producible by a residual factor vector can vary from the "optimistic surface" $O\hat{q}_1\hat{q}_3\bar{q}_2$ to the "pessimistic surface" $O\hat{q}_1\bar{q}_2$; and the project evaluator cannot predict what will happen and therefore which opportunity cost will prevail.[19]

Finally, note that the indeterminacy of the marginal variational shadow prices in this case renders somewhat academic the question whether their use outside the diversification cone would understate or overstate the true opportunity cost of factors withdrawn for project use.

D. *Case 5: Factors Outnumber Goods, No Distortion*

We now turn to the case where factors outnumber goods and there is no distortion. As one would expect from the well-known work of Paul Samuelson (1953) and other writers in the theory of international trade, even this distortion-free world will present problems, because in general primary factor prices will vary with the factor endowments so that the withdrawal of factors for project use will generally imply varying factor prices and hence the absence of a set of stationary shadow factor prices, except for a negligible set of residual factor vectors. More precisely, we will argue the following propositions concerning the implications of the case where factors outnumber goods:[20]

PROPOSITION 1: *Marginal variational shadow factor prices will be stationary if and only if the residual factor proportions belong to that negligible set of all possible propor-*

[18]The cone q_1Oq_2 is no longer a Rybczynski cone since shadow factor prices are not stationary within it any more because of the distortion.

[19]On the other hand, note that the uniqueness of opportunity cost will return once we leave the diversification cone.

[20]The formal proofs are to be found in the unpublished Appendix I in our working paper.

tions[21] *where the residual factors can be absorbed into industries at preproject factor intensities.*

PROPOSITION 2: *If the condition in Proposition 1 is violated, the Law of Variable Proportions will cause the opportunity cost for the project to differ from the marginal variational value of the factors withdrawn for use in the project, and to exceed the latter by a secondary cost.*

PROPOSITION 3: *This secondary cost will bear a proportional relation to the marginal variational value of the residual factors: a proportion rising steadily from zero at a nondecreasing rate as the residual factor proportions deviate progressively from the "negligible set" cited in Proposition 1.*

PROPOSITION 4: *There is a continuum of diversification subcones, each of them polyhedral in form and corresponding to a unique domestic factor-price vector. If that domestic factor-price vector deviates from the preproject factor-price vector, then all residual factor vectors in that subcone will be overvalued under (preproject) marginal variational valuation by the same proportion. Each diversification subcone is a Rybczynski cone (defined on some domestic factor-price vector) and the totality of all such diversification cones, not necessarily convex, forms the McKenzie-Chipman diversification cone.*

Propositions 1–3 can be visualized immediately by reference to a two-factor, one-good model. In Figure 4A, if the residual factor vector lies (as at A) on the ray OR from the origin, which is the ray on which the preproject factor endowment ratio lies, clearly there will be no project-induced change in the factor prices. For projects that leave the residual factor-endowment ratios elsewhere (as at B and C, for example, on the rays OB and OC), on the other hand, there will be corresponding changes in factor prices, and

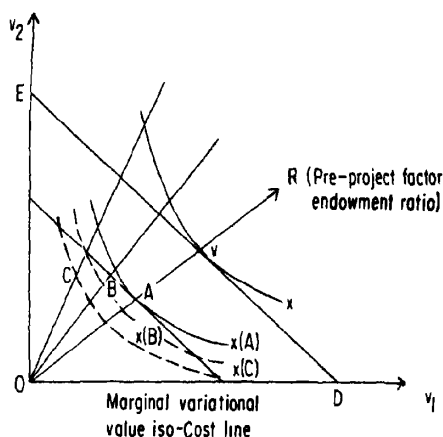



FIGURE 4A

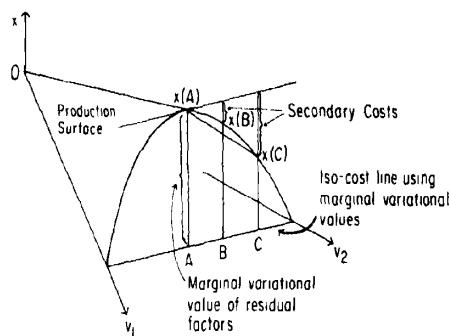



FIGURE 4B

the secondary cost that results may be illustrated in the now familiar construction (see Figure 3A) in Figure 4B. Note in particular that the values of the outputs for the isoquants passing through A, B, C and indeed all points along the ABC line in Figure 4A can be plotted as the concave curve in Figure 4B and this at once illustrates Proposition 3 above.[22] At the same time, note that the residual factor proportions must lie on the ray OR if stationarity of shadow factor prices is to be maintained, thus clearly demonstrating the "negligible" character of such stationarity-

[21]This set is the lower-dimension cone of the nonnegatively weighted sums of the preproject factor proportions for various industries.

[22]Recall that the curve $x(A)$ $x(B)$ $x(C)$ in Figure 4B is actually the vertical cross section of the production surface along the line ABC in Figure 4A.

preserving projects in our two-dimensional world, as argued in Propositions 1 and 2 above.

Turn now to the three-factor, two-good world where Proposition 4 is also admissible and hence demonstrable. (Since the formal proof is in Appendix I of our working paper, a few words of intuitive explanation should suffice here.) Corresponding to the initial preproject factor endowments, assume that the factor prices are determined such that all goods are produced (i.e., "diversification" obtains). Now, if for a finite project the postproject residual factor-endowment vector can be accommodated by suitable reweighting of the initial factor-proportions vectors, this is fine and stationarity obtains: though, as stated in Proposition 1, this set of possibilities is negligible. Suppose now that the residual factor-endowment vector requires new factor prices and hence new factor-proportion vectors in the production of the different goods. We can then immediately see both that there is now secondary cost and that, assuming continuing diversification, generally speaking there will be again a negligible set of factor-withdrawal possibilities at which the new factor prices and proportions, and hence the proportionate secondary cost vis-à-vis the old (marginal variational) factor prices, will remain unchanged. Therefore, clearly Proposition 4 is intuitively established.

E. *Case 6: Factors Outnumber Goods, with Distortion*

Here, as in Case 5, except for a negligible set of residual factor vectors, the marginal variational shadow factor prices may be nonstationary. The presence of the distortion, however, makes it impossible to assert that the use of marginal variational shadow factor prices outside of the diversification subcone to which they pertain will necessarily overstate or understate the true opportunity cost of the factors withdrawn for project use.

III. Concluding Remarks

Our analysis leads to many observations. First, as a quick perusal of the summary of Table 1 will show, the relative numbering of factors and goods is of significance. Project evaluation by shadow price computation is possible in the happy world of equal numbers of goods and factors, but lies between the Scylla of indeterminacy where goods exceed factors and the Charybdis of nonstationarity where factors exceed goods. The presence of distortions, in turn, is seen to be of significance in two ways: (i) unlike the distortion-free case, it creates nonuniqueness of shadow factor prices when goods outnumber factors; and (ii) while the distortion-free cases are characterized by secondary cost when marginal variational shadow factor prices are used outside of the zone of stationarity, the distortionary cases can be characterized instead by secondary gain as well.

Second, our analysis has clear applicability to the transfer problem, conceived *not* as a transfer of purchasing power, but rather as a transfer of factors of production as may be the case when reparations payments have to be made in barter (for example, Soviet Union transferring factories from Germany after World War II). By contrast with the purchasing power shift variety of transfers analyzed in the standard trade-theoretic literature (for example by Samuelson, 1952, 1954; Harry Johnson, 1956; and others), we must now contend with the possible existence of secondary costs (or gains, when distortions are present) even when the small-country assumption is made.

Third, our analysis also has applicability therefore to the theory of international factor mobility. Thus, the existing theoretical analyses, by Herbert Grubel and A. Scott, Albert Berry and Ronald Soligo, Johnson (1968), and Peter Kenen, of the welfare effects of brain drain on "those left behind," as reviewed and synthesized in Bhagwati and Carlos Rodriguez, relate to two-factor models with one or two products, with focus mainly on one-factor emigration. Our present analysis leads, however, to the following generalization: For a small country without distortions, and with each individual possessing the same factor endowment, any finite level of emigration of factors, singly or in combination, will harm (or have no effect on) those

left behind, if the residual factor vector falls outside (or is left inside) the preemigration Rybczynski cone. The harmful effect will almost always obtain when factors outnumber goods.

Fourth, for a small country without distortion, the welfare impact of finite factor increases (for example labor immigration, capital inflow) on those originally present is completely symmetrical. If every individual has an identical endowment, there is no effect *or* there is a beneficial effect (in the absence of distortions), depending upon whether the "augmented factor vector" is inside *or* outside the Rybczynski cone, and the latter possibility will almost always arise with factors outnumbering goods.

Fifth, the *locus classicus* of shadow prices is mathematical programming. It appears desirable to relate our findings with the programming framework, cross-referencing the economic assumptions, the programming formulation, and the implications for the shadow cost of a project. All these are tabulated in Table 2. Recall that, inside the Rybczynski cone, the Law of Variable Proportions is held at bay for the small-country, constant returns situation; hence stationarity of marginal variational shadow factor cost obtains.

TABLE 2—RELATIONSHIP OF ANALYSIS TO MATHEMATICAL PROGRAMMING

Set of Economic Assumptions	Programming Framework	Secondary Cost[a]
1) Large country, decreasing returns to scale	Non-linear programming	Ubiquitous
2) Small country, constant returns to scale, variable coefficients	Generalized linear programming of Dantzig-Wolfe[b]	Absent if residual factor vector stays inside the Rybczynski cone
3) Small country, fixed coefficients	Linear programming	Absent if residual factor vector stays inside the Rybczynski cone

[a]The tabulation is based upon the distortion-free cases. With distortion, the secondary cost can be a secondary gain.

[b]See and compare George Dantzig

REFERENCES

A. Berry and R. Soligo, "Some Welfare Aspects of International Migration," *J. Polit. Econ.*, Sept./Oct. 1969, *77*, 778–94.

T. Bertrand and F. Flatters, "Tariffs, Capital Accumulation and Immiserizing Growth," *J. Int. Econ.*, Nov. 1971, *1*, 453–60.

J. Bhagwati and T. N. Srinivasan, "The General Equilibrium Theory of Effective Protection and Resource Allocation," *J. Int. Econ.*, Aug. 1973, *3*, 259–82.

——— **and C. Rodriquez,** "Welfare-Theoretical Analyses of the Brain Drain," *J. Develop. Econ.*, Sept. 1974, *2*, 195–221; reprinted in Jagdish Bhagwati, ed., *The Brain Drain and Taxation: Theory and Empirical Analysis*, Amsterdam 1975.

———**, T. N. Srinivasan, and H. Wan, Jr.,** "Value Subtracted, Negative Shadow Prices of Factors in Project Evaluation, and Immiserizing Growth: Three Paradoxes in the Presence of Trade Distortions," *Econ. J.*, Mar. 1978, *88*, 121–25.

——— **and H. Wan, Jr.,** "Shadow Prices in Project Evaluation with and without Distortions and with Many Goods and Factors," econ. work. paper no. 205, M.I.T., July 1977.

R. Brecher, "Optimum Commercial Policy for a Minimum Wage Economy," *J. Int. Econ.*, May 1974, *4*, 139–50.

W. M. Corden, *Trade Policy and Economic Welfare*, London 1974.

George Dantzig, *Linear Programming and its Extensions*, Princeton 1963.

P. Diamond and J. Mirrlees, "Private Constant Returns and Public Shadow Prices," *Rev. Econ. Stud.*, Feb. 1976, *43*, 41–48.

R. Findlay and S. Wellisz, "Project Evaluation, Shadow Prices and Trade Policy," *J. Polit. Econ.*, June 1976, *84*, 543–52.

D. Gale and H. Nikaido, "The Jacobian Matrix and Global Univalence of Mappings," *Mathematische Annalen*, 1965, *159*.

H. Grubel and A. Scott, "The International Flow of Human Capital," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 268–74.

H. Johnson, "The Transfer Problem and Exchange Stability," *J. Polit. Econ.*, June 1956, *64*, 212–25.

———, "Some Economic Aspects of Brain Drain," *Pakistan Develop. Rev.*, No. 3, 1967, *7*, 379–411.

V. Joshi, "The Rationale and Relevance of the Little-Mirrlees Criterion," *Bull. Oxford Univ. Inst. Econ. Statist.*, Feb. 1972, *34*, 3–33.

P. Kenen, "Migration, the Terms of Trade and Welfare in the Source Country," in Jagdish Bhagwati et al., eds., *Trade, Balance of Payments and Growth*, Amsterdam 1971.

D. Lal, "Methods of Project Analysis: A Review," occas. paper no. 16, World Bank, Washington 1974.

Jacques Lesourne, *Cost-Benefit Analysis and Economic Theory*, Amsterdam 1975.

Ian Little and James Mirrlees, *Manual of Industrial Project Analysis in Developing Countries*, Vol. 2, Paris 1969.

A. Rudra, "Use of Shadow Prices in Project Evaluation," *Indian Econ. Rev.*, Apr. 1972, *7*, 1–15.

P. Samuelson, "The Transfer Problem and Transport Costs: The Terms of Trade when Impediments are Absent," *Econ. J.*, June 1952, *62*, 278–304.

———, "Prices of Factors and Goods in General Equilibrium," *Rev. Econ. Stud.*, No. 1, 1953, *21*, 1–20.

———, "The Transfer Problem and Transport Costs, II: Analysis of Effects of Trade Impediments," *Econ. J.*, June 1954, *64*, 264–89.

T. N. Srinivasan and J. Bhagwati, "Shadow Prices for Project Selection in the Presence of Distortions: Effective Rates of Protection and Domestic Resource Costs," *J. Polit. Econ.*, Feb. 1978, *86*, 97–116.

Social Security, the Supply of Labor, and Capital Accumulation

By SHENG CHENG HU*

The Social Security system has played an important role in the economic life of American families. It not only provides security for the elderly, but is a device for automatic stabilization, a method of income redistribution, as well as an important factor affecting capital accumulation and the supply of labor. The purpose of this paper is to analyze the long-run effects of the Social Security system in a growing economy. The model employed here extends and generalizes the neoclassical life cycle growth models of Peter A. Diamond and Paul A. Samuelson by explicitly allowing for an endogenous retirement decision and bequest motive. I consider an economy in which the population grows at a constant rate. Each individual lives for two periods. In the first period, he works full time, earning an income of w and paying a Social Security tax of T. In the second period, he works a fraction of time and then retires, receiving from the government a pension of z. He is to choose a consumption path, a retirement age, and an amount of bequest so as to maximize his lifetime utility. From these individual decisions and the assumption that the government budget is balanced each period, we derive the aggregate capital and labor supply functions and analyze the effects of changes in Social Security on capital accumulation and the equilibrium wage and interest rates. The present model is similar to that of Martin S. Feldstein in that retirement decisions are assumed endogenous. The main difference is that his is a partial equilibrium analysis while the model presented here is a general equilibrium model capable of analyzing long-run effects. I show that the short-run effects of Social Security depend primarily on the elasticities of the demand and supply of labor, and its long-run effects are influenced as well by the elasticities of savings and bequest. It is further shown that an appropriate Social Security system can increase the long-run well-being of the economy by causing the rate of return on capital to converge to the Golden Rule level. If, however, the tax and pension levels are tied to the individual working-retirement decisions, the system causes distortions in the labor market. Because of this distortional effect, the optimal Social Security does not necessarily lead to the Golden Rule.

I. The Model

A. *The Production Sector*

Consider a competitive economy in which the production function is given by

$$(1)\quad Q(t) = K(t)f(x(t)), \qquad f'(x) > 0, \quad f''(x) < 0$$

where $Q(t)$ is the level of output in period t, $K(t)$ the capital input, $x(t) \equiv L(t)/K(t)$ the labor-capital ratio, and $f(x)$ the output per unit of capital. Assume that $f'(x)$ satisfies the Inada conditions: $f'(\infty) = 0, f'(0) = \infty$.

The equilibrium wage rate w and rate of return on capital r are equal to the marginal products of labor and capital, respectively:

$$(2)\quad w(t) = f'(x(t))$$

$$(3)\quad r(t) = f(x(t)) - x(t)f'(x(t))$$

Combining equations (2) and (3) we can express the equilibrium relationship between r and w as

$$(4)\quad r = \varphi(w), \quad \varphi' = -x < 0, \qquad \varphi'' = -1/f''(x) > 0$$

This function defines the well-known factor-price frontier.

*Professor of economics, Purdue University. I wish to thank an anonymous referee and George Borts for numerous helpful comments and suggestions. This research is supported by a grant from the Social Security Administration.

B. *The Household Sector*

Assume that the total population of the economy $P(\mathrm{t})$ grows at a constant relative rate g. All individuals are alike, with the obvious exception of their ages. Each person born at (the beginning of period) t lives for two periods and is capable of providing one unit of labor per period. In the first period t he works full time, earning a wage income of $w(\mathrm{t})$ while paying a Social Security tax of $T(\mathrm{t})$. In addition he receives at the end of period t an inheritance of $h(\mathrm{t})$. In the second period $\mathrm{t}+1$, he works a fraction $(1-\alpha(\mathrm{t}+1))$ of the time, and then retires.[1] His earnings in the second period therefore consist of a net wage income of $(1-\alpha(\mathrm{t}+1))(w(\mathrm{t}+1)-T(\mathrm{t}+1))$ and a pension of $\alpha(\mathrm{t}+1)z$. For simplicity of the analysis, much of the paper will be devoted to the case where the pension level z is constant, but at the end of the paper I shall pay brief attention to the case where z is determined by a majority voting process.

Let the lifetime utility $U(\mathrm{t})$ of a representative individual born at t be dependent upon his consumption in both periods, the length of his retirement in the second period, and the amount of terminal wealth which he bequeaths to his children:

$$(5)\quad U(\mathrm{t})=U(c^1(\mathrm{t}),c^2(\mathrm{t}+1),\alpha(\mathrm{t}+1),\beta(\mathrm{t}+1))$$

where $c^i(\mathrm{t}+i-1)$ denotes his consumption in the ith period of his life (i.e., period $\mathrm{t}+i-1$), $\alpha(\mathrm{t}+1)$ the length of his retirement in the second period, and $\beta(\mathrm{t}+1)$ the amount of terminal wealth (or bequest). It is assumed that the lifetime utility function U is increasing and quasi concave with respect to all its arguments. For convenience of the analysis below, it is further assumed that U is additive so that the marginal rates of substitution among c^2, α, and β are independent of c^1. That is, U is assumed to be of the form

$$(5')\quad U=u(c^1)+v(c^2,\alpha,\beta)$$

The lifetime allocation problem for this person is then to choose α, β, c^1, and c^2 so as to maximize his utility function subject to his lifetime income constraint,[2] which can be written as

$$(6)\quad c^1(\mathrm{t})+\rho^e(\mathrm{t}+1)(c^2(\mathrm{t}+1)+\beta(\mathrm{t}+1))+\rho^e(\mathrm{t}+1)(w^e(\mathrm{t}+1)-T^e(\mathrm{t}+1)-z)\alpha(\mathrm{t})=y(\mathrm{t})$$

$$(7)\quad y(\mathrm{t})=w(\mathrm{t})-T(\mathrm{t})+\rho^e(\mathrm{t}+1)(w^e(\mathrm{t}+1)-T^e(\mathrm{t}+1))+h(\mathrm{t})$$

where $y(\mathrm{t})$ represents the "net" lifetime income,[3] $\rho(\mathrm{t}+1)=1/(1+r(\mathrm{t}+1))$ is the discount factor, and the superscript e over a variable denotes the expected value of that variable. It is seen from (6) that the prices of future consumption and bequest in terms of current consumption are $\rho^e(\mathrm{t}+1)$, while the price of lengthening retirement is $\rho^e(\mathrm{t}+1)(w^e(\mathrm{t}+1)-T^e(\mathrm{t}+1)-z)$.

The first-order conditions for maximum are[4]

$$(8)\quad \frac{\partial v}{\partial c^2}\Big/\frac{du}{dc^1}=\frac{\partial v}{\partial \beta}\Big/\frac{du}{dc^1}=\rho^e$$

$$(9)\quad \frac{\partial v}{\partial \alpha}\Big/\frac{\partial v}{\partial c^2}\geqq w^e-T^e-z$$

with equality if $\alpha<1$

[1] Although I consider only the case in which $\alpha \leq 1$ so that the individual retires in the second period, the analysis can be extended to the case where the individual retires completely in the second period and part of the time in the first period. But the result remains essentially unchanged.

[2] The case where the individual does not die until the end of the second period is considered here. In other words, the mortality rate of individuals aged 1 is zero, and that of individuals aged 2 is one. If the mortality rate of the younger generation is p, then the lifetime maximization problem becomes

$$\mathrm{Max}\ U=u(c^1)+(1-p)v^*(y^{2e},w^e(\mathrm{t}+1)-T^e(\mathrm{t}+1)-z,\rho^e)$$

where $v^*=\max v(c^2,\alpha,\beta)$ subject to $c^2+\beta+(w^e-T^e-z)\alpha=y^{2e}$ and $y^{2e}=w^e-T^e+(w-T+h-c^1)(1+r^e)$. The rest of the present analysis remains unchanged.

[3] His expected "disposable" lifetime income is

$$y_d(\mathrm{t})=w(\mathrm{t})-T(\mathrm{t})+\rho^e(\mathrm{t}+1)[(1-\alpha(\mathrm{t}+1))\cdot(w^e(\mathrm{t}+1)-T^e)(\mathrm{t}+1)+\alpha(\mathrm{t}+1)z]$$

[4] The second-order conditions are satisfied by the quasi-concavity assumption of U.

Equation (8) asserts that the marginal rate of substitution of present consumption for future consumption and for bequest must equal the discount factor. Equation (9) requires that the marginal rate of substitution between consumption for the second period and retirement be equal to the price of lengthening retirement (in terms of c^2). Although the marginal rate of substitution between c^2 and α is always positive, the price of lengthening retirement can be either positive or negative depending upon $w^e - T^e - z \lesseqgtr 0$. Consequently, it is possible for (9) to be held with inequality, in which case the individual retires completely in the second period and the present model reduces to that of Samuelson with bequest. The planned current and future consumption, planned length of retirement, and bequest level, obtained from (6)–(9), are dependent on the two prices and the lifetime income. In particular, we can write the current consumption function as[5]

$$
\begin{aligned}
(10)\quad c^1(t) &= \hat{C}^1(\rho^e(t+1), \\
&\quad \rho^e(t+1)(w^e(t+1) \\
&\quad - T^e(t+1) - z), y(t)) \\
&= c^1(w(t) - T(t), \\
&\quad w^e(t+1) - T^e(t+1), \\
&\quad r^e(t+1), h(t), z) \\
&\quad 1 > c_1^1 > 0,\ c_2^1 > 0,\ c_3^1 \lesseqgtr 0, \\
&\quad 1 > c_4^1 > 0,\ c_5^1 \lesseqgtr 0
\end{aligned}
$$

Assuming that c^1, c^2, α, and β are all normal goods and net substitutes, we can show upon using the Slutsky equation that a rise in the current net wage rate $(w(t) - T(t))$ increases current consumption by causing a rise in the lifetime income, but, because the income effects of an increase in $w(t) - T(t)$ on c^2, α, and β are all positive, the marginal propensity of current consumption with respect to the current net wage rate is less than one. A rise in the expected future net wage rate $(w^e(t+1) - T^e(t+1))$ increases current consumption not only because its income effect is positive, but also because it brings about substitution of c^1 for α. The substitution effect on c^1 of an increase in the expected interest rate $r^e(t+1)$ is negative while its income effect is positive; therefore, its total effect can be either positive or negative depending upon the relative strength of the substitution and income effects. An increase in the inheritance received raises current consumption, but by less than the increase in the inheritance. The effect on c^1 of an increase in the pension level is indeterminate, because its substitution effect is negative while its income effect is positive. By contrast, in the case where (9) is held with inequality so that $\alpha = 1$, the substitution effect is absent, and thus the increased pension level unequivocally raises current consumption.

The saving out of current income (excluding inheritance) is

$$
\begin{aligned}
(11)\quad s(t) &= w(t) - T(t) - c^1(t) \\
&= s(w(t) - T(t), \\
&\quad w^e(t+1) - T^e(t+1), \\
&\quad r^e(t+1), h(t), z) \\
&\quad 1 > s_1 > 0,\quad s_2 < 0,\quad s_3 \lesseqgtr 0, \\
&\quad 0 > s_4 > -1,\ s_5 \lesseqgtr 0
\end{aligned}
$$

Using (10) we easily obtain that $s(t)$ is increasing with respect to the current wage rate, and decreasing with respect to the expected future net wage rate and inheritance, but its response to a change in the expected interest rate or the pension level is indefinite.

By virtue of the additivity assumption, the levels of $c^2(t)$, $\alpha(t)$, and $\beta(t)$ chosen by an individual aged 2 in period t are determined by the following maximization problem:

$$\max\ v(c^2(t), \alpha(t), \beta(t))$$

subject to

$$(12)\quad c^2(t) + \beta(t) + (w(t) - T(t) - z)\alpha(t) = y^2(t),$$

[5]
$$
\begin{aligned}
c_1 &\equiv \frac{\partial c^1}{\partial(w - T)} = \frac{\partial C}{\partial y} > 0 \\
c_2 &\equiv \frac{\partial c^1}{\partial(w^e - T^e)} = \frac{\partial C}{\partial(w^e - T^e - z)\rho^e} \cdot \frac{\partial \rho^e(w^e - T^e - z)}{\partial(w^e - T^e)} + \frac{\partial C}{\partial y}\frac{\partial y}{\partial(w^e - T^e)} \\
&= \rho^e\Big[\frac{\partial C}{\partial(w^e - T^e - z)\rho^e}\Big|_{U\ \text{constant}} + (1 - \alpha)\frac{\partial C}{\partial y}\Big] > 0
\end{aligned}
$$

The rest of the derivatives can be derived in the same manner.

(13) $y^2(t) = w(t) - T(t) + (1 + r(t))(s(t-1) + h(t-1))$

where $s(t-1)$ is his savings in the preceding period and $h(t-1)$ is the inheritance he received. The optimality conditions are

(14) $$\frac{\partial v}{\partial \beta(t)} \Big/ \frac{\partial v}{\partial c^2(t)} = 1$$

(15) $$\frac{\partial v}{\partial \alpha(t)} \Big/ \frac{\partial v}{\partial c^2(t)} \geqq w(t) - T(t) - z,$$

with equality if $\alpha < 1$

Solving equations (12)–(15) we obtain

(16) $$\alpha(t) = \begin{cases} A(w(t) - T(t) - z, y^2(t)) \\ = \alpha(w(t) - T(t), r(t), k(t), z) \\ \text{with } \alpha_1 < 0,\ \alpha_2 > 0,\ \alpha_3 > 0,\ \alpha_4 > 0, \quad \text{if } w > \underline{w} \\ 1 \quad \text{if } w \leq \underline{w} \end{cases}$$

(17) $$\beta(t) = B(w(t) - T(t) - z, y^2(t)) = \beta(w(t) - T(t), r(t), k(t), z)$$
$$\beta_1 > 0,\ \beta_2 > 0,\ \beta_3 > 0,\ \beta_4 \gtrless 0$$

where $k(t) = s(t-1) + h(t-1)$ is the initial wealth held by this individual and $\underline{w} = \underline{w}(T(t), k(t), z)$ is the minimum value of w for which (15) is held with equality. Clearly, unless $w(t)$, $T(t)$, and $r(t)$ are what he expected in the previous period, the levels of $c^2(t)$, $\alpha(t)$, and $\beta(t)$ determined by this maximization problem are not equal to the levels he desired when he was aged 1. Under the usual assumption that the substitution effect dominates the income effect, the length of retirement determined by equation (16) is increasing when there is a decrease in the net wage rate and the individual retires completely in period 2 when the wage rate is below $\underline{w}$. The length of retirement also increases when there is an increase in the initial wealth, the pension level, or the interest rate. Likewise, the desired amount of bequest determined by (17) increases with respect to the net wage rate, the interest rate, or the initial wealth, but its response to changes in the pension level is uncertain. Let us assume that the bequest is evenly distributed among the younger generation, then

(18) $$h(t-1) = \beta(t-1)/(1+g)$$

II. The Short-Run Equilibrium

A. The Labor Market Equilibrium

The total supply of labor $L(t)$ in period t consists of the supply by the two generations. Since youngsters work full time while elders retire 100 $\alpha(t)$ percent of the time, we have

(19) $$L(t) = P(t)\,[1 - \theta + \theta(1 - \alpha(t))] = P(t)l(t)$$

(19′) $$l(t) = 1 - \theta\alpha(t)$$

where $\theta = 1/(2+g)$ is the proportion of the population who are aged 2, and $l(t)$ is the per capita aggregate supply of labor. Upon substituting (5) and (16) into (19′), we can rewrite the per capita labor supply function for $w \geq \underline{w}$ as

(20) $$l(t) = 1 - \theta\alpha(w(t) - T(t), \varphi(w(t)), k(t), z) = l(w(t), T(t), k(t), z)$$

For later use, define

$$\epsilon_w \equiv \partial \log l/\partial \log w = -(\alpha_1 + \alpha_2\varphi'(w))\theta w/l > 0$$
$$\epsilon_T \equiv \partial \log l/\partial \log T = \theta T\alpha_1/l < 0$$
$$\epsilon_k \equiv \partial \log l/\partial \log k = -\theta k\alpha_3/l < 0$$
$$\epsilon_z \equiv \partial \log l/\partial \log z = -\theta z\alpha_4/l < 0$$

The relationship between the per capita labor supply and the current wage rate, as determined by (20), is positive. However, such a relationship is not the usual labor supply curve because we have imposed here the equilibrium relationship between the interest rate and the wage rate, which is not the case of the usual labor supply curve. For convenience, let us call relation (20) the pseudo labor supply function. An interesting feature of this pseudo supply curve is that it is affected inversely by a change in the wealth possessed by the older generation and therefore the capital stock of the economy. It also shifts leftward when there is an increase in the Social Security tax or the pension level. As noted in the preceding section, if $w(t) \leq \underline{w}$ then the older generation retire completely and (20) becomes

(20′) $$l(t) = 1 - \theta \text{ for } w(t) \leq \underline{w}$$

In this case $\epsilon_w = \epsilon_z = \epsilon_T = \epsilon_k = 0$.

The supply of capital is provided solely by the older generation out of their past savings and the inheritance they received at the beginning of the period.

$$(21)\quad K(t) = P(t)\theta(s(t-1) + h(t-1)) \equiv P(t)\theta k(t)$$

This is exogenous within the period.

The equilibrium in the labor market requires that the per capita demand and supply of labor be equal. The per capita supply of labor $l(t)$ is given by (20). The per capita demand for labor $l^d(t)$ is determined by equating the marginal product of labor to the real wage rate $w(t) = f'(x(t))$, where $x(t) = L^d(t)/K(t) = l^d(t)/(\theta k(t))$. Therefore,

$$(22)\qquad w(t) = f'(l^d(t)/(\theta k(t)))$$

This equation defines the demand for labor as a decreasing function of the wage rate and an increasing function of the per capita capital stock (θk).

Combining (20) and (22) we can write the labor market equilibrium condition as

$$(23)\qquad w(t) = f'\left(\frac{l(w(t), T(t), k(t), z)}{\theta k(t)}\right)$$

Since the pseudo labor supply curve is upward sloped and the labor demand curve downward sloped, this equilibrium condition determines a unique short-run equilibrium wage rate. This equilibrium wage rate rises when there is an increase in the Social Security tax or the pension level, causing a leftward shift in the pseudo labor supply curve. It also rises when there is an increase in the capital stock, which leads to a leftward shift in the labor supply curve as well as a rightward shift in the demand curve.

B. *The Role of the Government*

Without further complications and great loss of generality, let us assume that the only role played by the government is to provide Social Security. Moreover, it finances Social Security by a "pay-as-you-go" system under which the government does not accumulate capital and it merely transfers payments by workers to retirees. The total government expenditure each period is then $G(t) = P(t)\theta\alpha(t)z = P(t)(1 - l(t))z$, and its revenue is $R(t) = P(t)l(t)T(t)$.

Because the government does not accumulate capital, the budget equilibrium condition is given by $R(t) = G(t)$, or

$$(24)\quad z = (T(t) + z)l(w(t), T(t), k(t), z)$$

Assuming that $\xi \equiv T + (T + z)\epsilon_T > 0$, we can solve this equation for a unique tax level for any given $w(t)$, $k(t)$, and z,

$$(25)\qquad T(t) = F(w(t), k(t), z)$$

with $\partial log\ F/\partial log\ w = -(T + z)\epsilon_w/\xi \leq 0$, $\partial log\ F/\partial log\ k = -(T + z)\epsilon_k/\xi \geq 0$, and $\partial log\ F/\partial log\ z = (T - (T + z)\epsilon_z)/\xi \geq 1$, where the equalities hold if $w \leq \underline{w}$. For $w > \underline{w}$, the tax level falls when there is a rise in the wage rate, causing a delay in retirement by the older generation and thereby a reduction in the government spending on Social Security; but it rises when there is an increase in the pension level or the capital stock, which brings about earlier retirement. For $w \leq \underline{w}$, the tax level increases proportionately with z, but is invariant to changes in w or k.

C. *The Overall Equilibrium*

The overall short-run equilibrium of the economy is attained if both the labor market is cleared and the government budget is balanced. Thus, the overall short-run equilibrium wage rate and tax level, labeled $w^+(t)$ and $T^+(t)$, respectively, can be determined by substituting (25) into (23). They are functions of z and $k(t)$,

$$(26)\qquad w^+(t) = w^+(k(t), z)$$

$$(27)\qquad T^+(t) = T^+(k(t), z)$$

The equilibrium wage rate is shown in Figure 1 by the intersection of the l^d and the l^c curves. The downward-sloping l^d curve denotes the demand for labor. The l^c curve is the "compensated supply of labor curve," which depicts the level of labor supply at each w when any change in w is compensated by a corresponding change in the Social Security tax so as to maintain a balanced government

a. Effects of An Increase in the Pension Level

b. Effects of An Increase in the Capital Stock

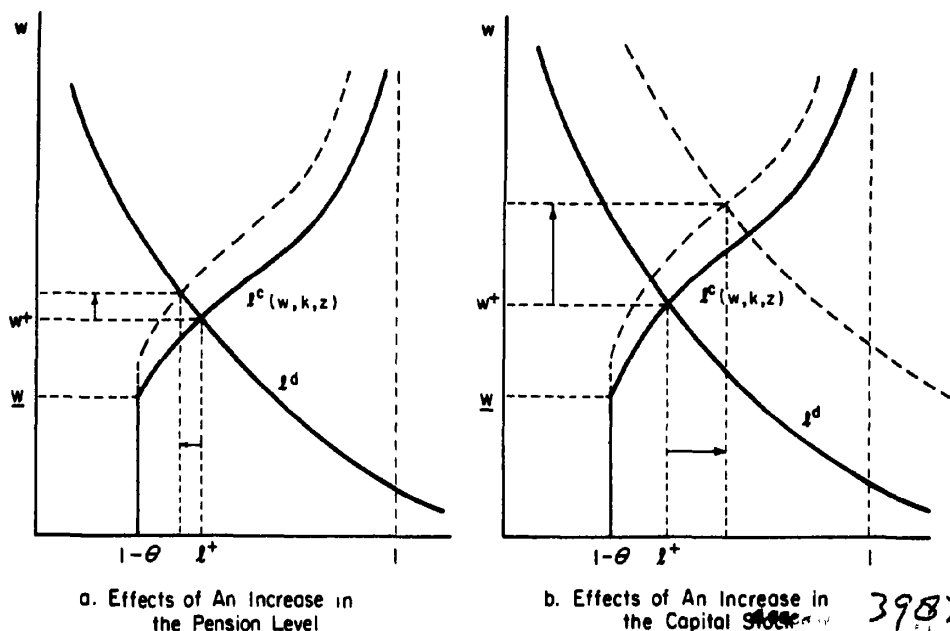



FIGURE 1. SHORT-RUN EQUILIBRIUM

budget; i.e., l^c is obtained by substituting (25) into (20) and (20′).

$$(28)\quad l^c = \begin{cases} l(w, F(w, k, z), k, z) \\ \quad \equiv l^c(w, k, z), \text{ for } w > \underline{w} \\ 1 - \theta \qquad\qquad \text{for } w \leq \underline{w} \end{cases}$$

It can be shown that the l^c curve is upward sloped for $w > \underline{w}$ and becomes vertical for $w \leq \underline{w}$.[6] It is seen from Figure 1 that depending upon the position of the l^d curve, the short-run equilibrium can take place at either the vertical or the upward-sloping region of the l^c curve.

Consider now the effects of changes in z and $k(t)$. Upon differentiating (23) and (24) jointly with respect to z, we obtain that

$$(29)\quad \frac{\partial\, log\, w^+(t)}{\partial\, log\, z} = -\frac{1}{\eta}\frac{\partial\, log\, l^+(t)}{\partial\, log\, z} = -\frac{T}{\gamma}(\epsilon_z + \epsilon_T) \geq 0$$

$$(30)\quad \frac{\partial\, log\, T^+(t)}{\partial\, log\, z} = -\frac{1}{\gamma} \cdot [(\eta + \epsilon_w)T - \eta\epsilon_z(T + z)] \geq 1$$

where $\eta \equiv -\partial log\ l^d/\partial log\ w > 0$ is the elasticity of the demand for labor and $\gamma \equiv \epsilon_w T + \xi\eta > 0$. The effect of an increase in the pension level is to shift the l^c curve to the left, which in turn brings about an increase in the equilibrium wage rate and a fall in the equilibrium employment level. It is obvious from (24) that when the equilibrium level of employment falls, the equilibrium tax level rises by a greater proportion than the increase in z. If $w^+(t) \leq \underline{w}$, then because $\epsilon_z = \epsilon_T = 0$, $w^+(t)$ is independent of a small change in z, while $T^+(t)$ increases proportionately with respect to z.

Differentiating (23) and (24) with respect to $k(t)$ yields

$$(31)\quad \frac{\partial\, log\, w^+(t)}{\partial\, log\, k(t)} = \frac{1}{\gamma}(\xi - \epsilon_k T) > 0$$

$$(32)\quad \frac{\partial\, log\, T^+(t)}{\partial\, log\, k(t)} = -\frac{T + z}{\gamma} \cdot (\eta\epsilon_k + \epsilon_w) \gtreqless 0 \text{ as } \eta\epsilon_k + \epsilon_w \lesseqgtr 0$$

[6]Upon differentiating l^c totally, we have

$$dlog\ l^c = (1/\xi)\{T\epsilon_w\, dlog\ w + T(\epsilon_z + \epsilon_T)\, dlog\ z + T\epsilon_k\, dlog\ k\}$$

Thus l^c increases with w but decreases with z and k.

An increase in capital shifts the l^d curve to the right and the l^s curve to the left. Therefore, its effect is to raise the short-run wage rate. Whether it results in a rise in the equilibrium level of employment and therefore a fall in the tax rate depends on whether the rightward shift in the l^d curve exceeds the leftward shift in the l^s curve. Equation (32) shows that the more elastic the labor demand, or the more inelastic the labor supply, the more likely the equilibrium employment of labor will rise and the Social Security tax will fall.

III. The Long-Run Equilibrium

A. *Comparative Analysis*

Although both $s(t-1)$ and $h(t-1) = \beta(t-1)/(1+g)$ are constant in the short run, they are variable over time and are determined by (11) and (17), respectively. Assuming static expectations, i.e., $w^e(t+1) = w(t)$, $r^e(t+1) = r(t)$, and $T^e(t+1) = T(t)$, we have

$$(11') \quad s(t) = s(w^+(t) - T^+(t), w^+(t) - T^+(t), \varphi(w^+(t)), h(t), z)$$

$$(17') \quad h(t) = \beta(w^+(t) - T^+(t), \varphi(w^+(t)), k(t), z)/(1+g)$$

where $w^+(t)$ and $T^+(t)$ are given by (26) and (27), respectively, and $k(t) = s(t-1) + h(t-1)$. We are particularly interested in the long-run equilibrium at which $s(t)$ and $h(t)$ are constant. Setting $s(t) = s^*$, $h(t) = h^*$, $w(t) = w^*$ and $T(t) = T^*$ in (11′), (17′), (23), and (24), we can determine the long-run equilibrium values of these endogenous variables:

$$(11'') \quad s^* = s(w^* - T^*, w^* - T^*, \varphi(w^*), h^*, z)$$

$$(17'') \quad \beta^* = \beta(w^* - T^*, \varphi(w^*), k^*, z)$$

$$(23') \quad w^* = f'(l(w^*, T^*, k, z)/(\theta k^*))$$

$$(24') \quad z = (z + T^*)l(w^*, T^*, k, z)$$

$$\text{or} \quad T^* = F(w^*, k^*, z)$$

Differentiating these four equations totally and making appropriate rearrangements we obtain

$$(33) \quad \frac{d\log k^*}{d\log z} = \frac{1}{\lambda}\{\gamma(\sigma_z + \sigma_T) - (\epsilon_z + \epsilon_T)[T\sigma_w + (T+z)\eta\sigma_T]\}$$

$$(34) \quad \frac{d\log l^*}{d\log z} = \frac{T}{\lambda}\{(\eta\epsilon_k + \epsilon_w)(\sigma_z + \sigma_T) - (\epsilon_z + \epsilon_T)[\sigma_w - \eta(1 - \sigma_k]\}$$

$$(35) \quad \frac{d\log w^*}{d\log z} = \frac{1}{\lambda}\{(\xi - \epsilon_k T)(\sigma_z + \sigma_T) - (\epsilon_z + \epsilon_T)[T(1 - \sigma_k) + (T+z)\sigma_T]\}$$

where $\lambda = \gamma(1 - \sigma_k) + (\eta\epsilon_k + \epsilon_w)(T+z)\sigma_T - (\xi - \epsilon_k T)\sigma_w$, which is assumed to be positive in consideration of the stability of the long-run equilibrium.

Let us define

$$\sigma_w = [s_1 + s_2 + s_3\varphi' + (\beta_1 + \beta_2\varphi')/(1+g)](w/k)$$

$$\sigma_T = -[s_1 + s_2 + \beta_1/(1+g)](T/k)$$

$$\sigma_z = [s_5 + \beta_4/(1+g)](z/k)$$

$$\sigma_k = \beta_3[s_4 + 1/(1+g)]$$

They are the elasticities of the supply of capital with respect to the wage rate, Social Security tax, pension level, and the capital stock existing in the previous period, respectively. The sign of σ_w is generally indeterminate; it is more likely to be positive the lower are the labor-capital ratio (i.e., $-\varphi'$) and the marginal propensity to save with respect to the future net wage rate. The sign σ_k is nonnegative and equals zero when there is no bequest motive; σ_T tends to be negative unless the marginal propensity to save is high with respect to the future net wage rate; σ_z is negative if the income effects of an increase in z on c^1 and β exceed its substitution effects, but is positive in the opposite case. When the equilibrium wage rate is less than $\underline{w}$ so that $\alpha = 1$, both σ_T and σ_z are unequivocally negative.

Equation (33) shows the effect of a change in pension benefits on the long-run equilibrium capital stock. This effect comes through the usual substitution and income effects. First, it induces substitution among c^1, c^2, α, and β by bringing about a change in their

relative prices both directly and indirectly (through its impact on the short-run equilibrium wage rate as well as tax level). Second, it results in a change in the level of the individual lifetime income as well as its distribution between the two periods, which, given the consumption stream, also contributes to a further change in the savings level. Since $\gamma > 0$ and $\epsilon_z + \epsilon_T \leq 0$, the total effect of an increase in pension benefits is to decrease the long-run equilibrium level of per capita capital stock (θk) if $\sigma_z + \sigma_T \leq 0$ and $T\sigma_w + (T + z)\eta\sigma_T \leq 0$. These conditions are satisfied if both σ_z and σ_T are negative and the elasticity of labor demand is high. In the converse case where $\sigma_z + \sigma_T \geq 0$ and $T\sigma_w + (T + z)\eta\sigma_T \geq 0$, the total effect is positive. In the special case considered by Samuelson where $\alpha = 1$ and $\epsilon_z + \epsilon_T = 0$, equation (33) reduces to

$$dlog\ k^*/dlog\ z = \eta(\sigma_z + \sigma_T)/(\eta + \sigma_w)$$

and $\sigma_z + \sigma_T < 0$. Therefore, the effect is unquestionably negative. But contrary to Samuelson, the elasticity of k^* with respect to z is not necessarily unitary. The difference results from the fact that Samuelson assumes a "fully funded social security," under which the government not only accumulates capital but pays an interest of r on Social Security. The present model, on the other hand, assumes a pay-as-you-go system in which the government does not accumulate capital and it pays an interest rate of g instead of r on Social Security. Therefore, the Social Security system not only alters the relative cost of retirement but produces an income effect based on the difference between r and g. My result in (33) is also different from Feldstein in that (i) he considers only the short-run partial equlibrium effect of Social Security, ignoring the impact of the change in pension benefits on the equilibrium wage rate, and (ii) he assumes that the Social Security tax paid in period 1 earns an interest rate of r, and therefore the income effect of Social Security mentioned above does not appear in his model either.

Let us turn next to the effects of increased pension benefits on the long-run equilibrium level of employment and wage rate. As noted above, the increase in pension benefits directly shifts the compensated labor supply curve to the left. But over the long run it also causes a change in the capital stock, which in turn affects the compensated labor supply curve inversely and the demand for labor curve in a positive direction. Therefore, while its short-run effect is to induce earlier retirement of the older generation and thereby to increase the wage rate, its long-run effect can be reversed, depending upon the direction and strength of this secondary effect. Considering the standard case where $\sigma_z + \sigma_T \leq 0$, we see from equation (34) that its long-run effect is to decrease the retirement age of the older generation if $\eta\epsilon_k + \epsilon_w \geq 0 \geq \sigma_w - \eta(1 - \sigma_k)$ while the converse is the case if $\eta\epsilon_k + \epsilon_w \leq 0 \leq \sigma_w - \eta(1 - \sigma_k)$. Likewise, equation (35) shows that the increased pension level lowers the equilibrium wage rate if $T(1 - \sigma_k) + (T + z)\sigma_T \leq 0$, or $\epsilon_z + \epsilon_T$ is sufficiently low. In the special case where $\alpha = 1$, because $\epsilon_z = \epsilon_T = 0$, (35) reduces to

$$dlog\ w^*/dlog\ z = (\sigma_z + \sigma_T)/(\eta + \sigma_w)$$

which is unequivocally negative.

From (24) we have that

$$dlog\ T^*/dlog\ z = 1 - (dlog\ l^*/dlog\ z)(T + z)/T$$

Therefore, if the increased pension level leads to a decrease in the long-run equilibrium level of employment or retirement age, it also leads to a rise in the Social Security tax more than proportionately.

B. *Welfare Implications*

A good measure of the long-run incidence of Social Security is in terms of its effect on the long-run equilibrium level of lifetime utility enjoyed by a representative individual. The effect of an increase in the pension level is twofold. First, it causes the marginal rate of substitution between consumption and retirement to diverge from the wage rate and therefore the marginal product of labor. This distortion of the labor market represents the static inefficiency, which tends to reduce the welfare of the individual. But, on the other hand, it causes intergenerational redistribution of income and a change in the long-run equilibrium level of capital, which increases

the welfare of the individual if it causes the rate of return on capital to converge to the Golden Rule level. Whether the total effect is positive depends on whether the increased dynamic efficiency exceeds the static inefficiency it brings about.

Differentiating (5) with respect to z and making use of (8) and (9) we have

$$\frac{dU^*}{dz} = \frac{\partial U}{\partial c^1}\left[\frac{dc^1}{dz} + \rho\frac{dc^2}{dz} + \rho(w - T - z)\frac{d\alpha}{dz} + \rho\frac{d\beta}{dz}\right]$$

Using the budget constraint (6), we can further reduce this expression to

$$= \frac{\partial U}{\partial c^1}\left[\frac{dy}{dz} - (c^2 + \beta)\frac{d\rho}{dz} - \alpha\frac{d\rho(w - T - z)}{dz}\right]$$

Expanding the terms inside the brackets yields

$$(36)\quad = \frac{\partial U}{\partial c^1}\left[(r^* - g)\rho^*\frac{dw^*}{dz} - (1 + \rho^* - \alpha\rho^*)\frac{dT^*}{dz} + \alpha\rho^* + \frac{1}{1+g}\frac{d\beta^*}{dz}\right]$$

Setting (36) equal to zero, we obtain the following condition with which to determine the optimal pension level:

$$(37)\quad (r^* - g)\rho^*\frac{dw^*}{dz} - (1 + \rho^* - \alpha\rho^*) \cdot \frac{dT^*}{dz} + \frac{1}{1+g}\frac{d\beta^*}{dz} = -\alpha\rho^*$$

Because

$$(1 + \rho - \alpha\rho)(T/z) - \alpha\rho$$
$$= (1 + \rho - \alpha\rho(T + z)/T)(T/z)$$
$$= (1 + \rho - \rho/\theta)(T/z) \quad \text{(by (24) and (19'))}$$
$$= (r - g)(T/z)\rho$$

equation (37) can be further reduced to

$$(38)\quad s^*(r^* - g) \cdot \left(\frac{dw^*}{dz} - \frac{T}{z}\right) - (1 + \rho^* - \alpha\rho^*) \cdot \left(\frac{d\log T^*}{d\log z} - 1\right)\left(\frac{T}{z}\right) + \frac{1}{1+\rho}\frac{d\beta^*}{dz} = 0$$

Clearly, if there is no bequest motive and the retirement decision is exogenous, the second and third terms in (38) disappear. Then the optimal level of z obeys the Golden Rule of capital accumulation: it equates the rate of return on capital (or equivalently the interest rate) to the population growth rate.

In the converse case where the bequest motive is present or the retirement decision is endogenous, the Golden Rule no longer holds. In particular, if $d \log l^*/d \log z < 0$ and $d \log \beta^*/d \log z < 0$, the optimal level of z obtains at a rate of interest lower than the growth rate of population. This can be attributed to two reasons. First, with the introduction of a bequest motive, the capital owned by the older generation earns not only an interest rate of r, but, when bequeathed to their children, also yields a nonpecuniary return, which is equal to $(\partial U/\partial\beta)/(\partial U/\partial c^1)$.[7] Thus, equality of the rate of return on capital to the growth rate of population implies that the equilibrium interest rate is below the population growth rate. Furthermore, as mentioned above, since both taxes and benefits are tied to the endogenous retirement decisions, the increased dynamic efficiency brought about by the Social Security system is accompanied by an increase in the static inefficiency. Therefore, the optimum, at which the increased dynamic efficiency is offset by the decreased static efficiency, does not equate the rate of return on capital to the population growth rate unless the latter is equal to zero.

C. *Majority Voting Equilibrium*

The optimal level of pension benefits having been determined, a natural question that arises is whether such an optimal level can be achieved in a democratic society

[7] See Mordecai Kurz for a similar argument.

through the majority voting process. In a recent paper, Edgar Browning has argued that in an economy in which there are three generations, the majority voting equilibrium is determined by the votes cast by the middle-aged group, and the level so determined exceeds the social optimum, at which the lifetime utility of a representative individual is maximized. Elsewhere I have considered an economy with heterogeneous income groups and have shown that the majority voting equilibrium level of Social Security, determined by the votes of the median income group, in general does not maximize the social utility function defined as the sum of individual lifetime utilities. The present model enables us to reconsider this problem: Here, because all individuals are alike and there are only two age groups, the majority voting equilibrium level of z is one which is desired by the representative youngster and therefore maximizes his lifetime utility. Consequently, the questions raised by Browning and myself do not arise. The problem now becomes whether the majority voting equilibrium converges to the optimal long-run equilibrium, determined by (38).

How each youngster votes clearly depends on his expectation regarding the implication of the change in z for the tax rate. Let us assume that he knows that the tax rates are given by (25), and that he does not expect any change in z to affect the wage rate. This implies that the level of z he desires is the solution to the following maximization problem:

$$\underset{\alpha,\beta,c^1,c^2,z}{\text{Maximize}}\ U(c^1, c^2, \alpha, \beta)$$

subject to equations (6) and (25). It satisfies (8), (9), and the following condition:

$$(39)\quad (1+\rho-\alpha\rho)\frac{\partial F}{\partial z} - \rho\alpha - \frac{1}{1+g}\frac{\partial \beta}{\partial z} = 0$$

Comparing (39) and (38) it follows that the majority voting equilibrium in general does not achieve the long-run optimum. And there is no reason to expect that it should be biased in any particular direction.

IV. Concluding Remarks

I have shown that the short-run effect of an increase in the pension level is to increase the gross wage rate and reduce the equilibrium level of employment or retirement age, but its long-run effect could be reversed. Although the pension level per unit of time has been assumed to be constant, independent of previous taxes paid, it is eventually related to taxes and wage income through the government budget constraint. In deriving the optimal structure of Social Security, we have been concerned only with its long-run consequences. It is possible that the long-run gain in the social welfare that results from a change in z may involve sacrifices by some generations during the transition period. To obtain the true optimal path of z, it is necessary to specify a social welfare function based on the utility enjoyed by each generation. It should also be noted that in a democratic society where Social Security is determined by the majority voting process, such a level generally does not converge to the one which maximizes individual utility in the long-run equilibrium.

REFERENCES

E. K. Browning, "Why the Social Insurance is too Large in a Democracy," *Econ. Inquiry*, Sept. 1975, *13*, 373–88.

P. A. Diamond, "National Debt in a Neoclassical Growth Model," *Amer. Econ Rev.*, Dec. 1965, *55*, 1126–50.

M. S. Feldstein, "Social Security and Private Saving: International Evidence in an Extended Life-Cycle Model," in Martin Feldstein and Robert Inman, eds., *The Economics of Public Services*, London 1977, 174–205.

S. C. Hu, "On the Dynamic Behavior of the Consumer and the Optimal Provision of Social Security," *Rev. Econ. Stud.*, Oct. 1978, *45*, 437–45.

M. Kurz, "Optimal Economic Growth and Wealth Effects," *Int. Econ. Rev.*, Oct. 1968, *9*, 348–58.

P. A. Samuelson, "Optimal Social Security in a Life Cylce Growth Model," *Int. Econ. Rev.*, Oct. 1975, *16*, 539–44.

Informative Advertising and Welfare

By Y. KOTOWITZ AND F. MATHEWSON*

The literature on the welfare consequences of advertising[1] is generally contentious and involves arbitrary implicit assumptions about consumer behavior and tastes. At the one extreme, there are those who believe that advertising changes consumer tastes, so that a standard of value is lacking to judge its welfare effects (see, for example, Nicolas Kaldor). Much of this analysis therefore imposes an external (personal?) standard in judging any welfare effects (see, for example, John Kenneth Galbraith). At the other extreme, there are those (George Stigler; Ferguson; Phillip Nelson, 1974; Lester Telser) who claim that advertising is an efficient means of transmitting information and consequently is generally beneficial. Implicit is the assumption (not generally justified) that the information transmitted is truthful. However, even if the assumption of advertising as information is accepted, it is not obvious that monopolists supply the socially optimal amount of it.

Recently, Avinash Dixit and Victor Norman attempted to analyze the welfare consequences of advertising. However, their analysis is performed with arbitrary assumptions about the way in which advertising affects demand.[2] Moreover, the analysis is conducted in a comparative static framework, which is inappropriate to the analysis of advertising as information, where consumers are learning about products through time.

If the informational aspects of advertising are to be analyzed properly, it is necessary to incorporate the analysis of advertising within a diffusion process and to contrast alternative forms of informational channels. This paper analyzes the welfare implications of simple alternative diffusion processes in monopolistic markets. Two types of processes are considered: (a) diffusion of information through informative or disseminative advertising by the monopolist, and (b) diffusion of information through demonstration effects by buyers. We derive optimal seller policies with respect to price and advertising over time and contrast them with the socially optimal policies. The main points emerging from the analysis are that a monopolist, following an optimal dynamic policy for profit maximization, generally:

(a) Produces less information in the form of disseminative advertising than is socially optimal, where demonstration effects are absent.

(b) Charges a lower price where demonstration effects are important, relative to that charged in their absence. However, the diffusion rates generated by these policies still fall short of those which are socially optimal.

(c) May supply excessive amounts of disseminative advertising relative to the social optimum, where demonstration effects are present.

We define disseminative advertising as advertising which informs potential buyers about the existence, price, and characteristics of the products. These characteristics must be well-defined and readily and inexpensively verifiable prior to purchase. (This corresponds to search quality in Nelson's 1970 terminology.) In this case, the only viable role of advertising is to inform those potential customers who do not know about the product or about its price and characteristics. This type of advertising is only part of all advertising by sellers. Where the measurement of the

*Professor and associate professor, respectively, Institute for Policy Analysis, University of Toronto. We are grateful to the Department of Consumer and Corporate Affairs, Government of Canada, and the Canada Council for financial support for this work, and for the comments and suggestions received at the department of political economy, University College, London during a sabbatical year for one of the authors.

[1]For surveys of the literature see Richard Schmalensee, pp. 4–9, and James Ferguson.

[2]For example, they impose the condition that all advertising necessarily decreases demand elasticity. But this is *not* a universally accepted condition. For example, Ferguson argues that the contrary holds.

quality embodied in a product is difficult and/or expensive and may involve prolonged use, advertising may operate so as to change buyer perceptions of quality or product attributes, even when they have used the product. We label this type of advertising as "persuasive."[3]

Demonstration effects occur where potential consumers, who are not aware of the product or its characteristics, acquire such information from those who have already bought the product. In this case, existing consumers act as agents of the seller in disseminating information about the product. This may occur simply by observation, as in the case of a new model or brand of car or aluminum siding on a house, or through conscious advice and recommendation by satisfied customers. Such processes for the spread of information are important in the growing popularity of new products, and consequently, it is important to investigate their positive as well as normative aspects.

Section I sets out the models and analyzes separately the optimal dynamic policy of a profit-maximizing monopolist through advertising and demonstration effects. Section II analyzes the social welfare effects of each policy separately. Section III defines the optimal monopoly policy with both advertising and demonstration effects, and contrasts these monopoly policies with those that maximize social welfare.

I. Diffusion Policies of a Monopoly

We consider a profit-maximizing monopolist producing a single output that has, at each moment of time, a readily verifiable and unchanging level of quality. (The monopoly exists either because of fixed costs or due to an exclusive government franchise.) Out of a total population of S, only N ($\leq S$) consumers are aware of the existence of the good and only M ($\leq N$) consumers actually purchase the product. Define $n \equiv N/S$ as the proportion of the population informed and $m \equiv M/N$ as the proportion of those informed who buy the product. (Both m and n are assumed to be continuous variables.) As defined, n represents an extensive margin, and m represents an intensive margin.

In this model, each consumer who purchases the good buys a given quantity of the good (set equal to one) per time period. This assumption has no effect on the substantive results and greatly simplifies the analysis by identifying units of the commodity purchased with the consumers who purchase them.[4] We define the utility function for consumers as

$$U^i = \alpha(i)\zeta + Z^i \tag{1}$$

where $\alpha(i)$ is a scalar which represents the intensity of preference by consumer i, $\zeta = 1$ if the consumer buys the good, $\zeta = 0$ otherwise. ($\zeta = 0$ for uninformed consumers $(1 - n)$, and for informed consumers who choose not to buy the good $n(1 - m)$.)

Within n, the proportion of the population aware of the good's existence and price, we rank consumers over the unit interval according to their intensity of preference $\alpha(i)$. We assume that the resulting distribution is continuous and differentiable ($\alpha' < 0$). As well, we assume that the distribution of tastes is independent of n; by this we mean that newly informed consumers have the same taste distribution as those already informed.[5] Consumers maximize their utility subject to a budget constraint bounded by an equal per capita distribution of rent (labeled r) on a basic productive resource. If the price per unit of the monopolist's output is p, then we may

[3]Persuasive advertising is treated in a separate paper by the authors.

[4]If a consumer purchases more than one unit of the commodity, then that consumer is treated as many consumers. Associated with each fictional consumer is a unit of the good, valued at the marginal utility of that unit. Fictional consumers are then ordered according to declining marginal valuations.

[5]This assumes that those who potentially stand to benefit most from the information will not make it their business to acquire it first. However, this assumption is not unreasonable where the consumers are totally ignorant of the product. The assumption also implies that the monopolist is either ignorant of potential consumers' tastes or is unable to direct the information to those most likely to buy first. To the extent that these assumptions do not hold, m is likely to become a negative function of n with consequent complications in the analysis. The substantive qualitative results remain unaffected.

identify a marginal consumer (m) who is just indifferent between buying and not buying the product:[6]

$$\alpha(m) = p \tag{2}$$

If M is the number of buyers, the monopolist's revenues are

$$R = \alpha(m)\, M \tag{3}$$

Output costs are assumed to be linear in output:

$$C = cM \tag{4}$$

Consumers who are currently uninformed of the product's existence are assumed to become informed in one of two possible ways, either by disseminative advertising expenditures or through a demonstration effect that is larger the greater the proportion of the population currently using the product. These are two corresponding simple specifications of the rate of change of informed consumers:

$$DN = g(a)(S - N) - \beta N \tag{5}$$

$$DN = (\delta/n)(M/S)(S - N) - \beta N \tag{6}$$

where D represents the operator d/dt, a represents dollars of advertising expenditure per capita, and $\beta (0 \leq \beta \leq 1)$ represents the proportion of the population who forget during each period. John Gould pointed out that equation (5) (with an appropriately positive but diminishing marginal product of advertising, i.e., $0 < g(a) \leq 1$, $g' > 0$, $g'' < 0$) corresponds to Stigler's model of the spread of information.

Equation (6) is the analogue to (5), with information being generated through a demonstration effect. The model assumes that buyers of the product disseminate information about it in proportion to contacts between buyers and those who do not know about it. If contacts were random, the number of these contacts would be proportional to the product of the number of buyers and the proportion of those who are unaware of the product. However, the number and effectiveness of such contacts is likely to diminish as the proportion of those who know about the product rises, because social and informational contacts occur with greater frequency within social networks than among them. Thus when n is small, information is transmitted readily within social networks. As n rises, knowledge must be increasingly transmitted between networks, as well as within relatively less efficient networks, so that the number and effectiveness of contacts are reduced. This effect is captured by the term $\delta/n \equiv \delta/(N/S)$, where δ is a constant and $(\delta/n)(M/S) \equiv \delta m \leq 1$.[7]

If the population grows at an exogenous positive rate of η (i.e., $S = S_0 e^{\eta t}$), all variables may be measured in per capita terms and we may define two monopoly problems which correspond to each of the specifications for the spread of information. These may be formally stated as[8]

Advertising Problem:

$$\max_{m,a} \; J^a = \int_0^\infty e^{-(\rho-\eta)t}\left[(p-c)mn - a\right] dt$$

subject to $Dn = g(a)\,(1-n) - \xi n$

Demonstration Problem:

$$\max_{m} \; J^m = \int_0^\infty e^{-(\rho-\eta)t}\,(p-c)mn\, dt$$

subject to $Dn = \delta m(1 - n) - \xi n$

where $\xi \equiv \eta + \beta$, $\rho > \eta$ and S_0, a constant, is suppressed.

[6]In the case where consumers purchase more than one unit of the good, $\alpha(m)$ represents the valuation of the marginal fictional consumer. (See fn. 4 above.)

[7]Our equation (6) is similar to S.A. Ozga's diffusion model, as represented in Gould's second diffusion model and to Stephen Glaister's model. However, the strength of contacts in these models is assumed not to decline with n. Consequently these models yield logistic-type relationships with an initial threshold effect. Thus information campaigns which cannot achieve the minimum threshold are doomed to failure. We eschew such specifications for two reasons: first, there is considerable evidence to suggest that contacts are not random and that information is transmitted more readily within networks (see, for example, Irwin Bernhardt and Kenneth Mackenzie; James Coleman, Elihu Katz, and Herbert Menzel; Katz, Martin Levin, and Herbert Hamilton); and second, the procedure we have chosen is much simpler to analyze and is sufficient to illustrate our welfare propositions.

[8]While the monopoly policy variables are price and advertising, it is convenient to set the problem in terms of the proportion of buyers (m), and advertising. $(\partial p/\partial m) < 0$, i.e., a lower price leads to greater purchases.

FIGURE 1

Both of these simple dynamic problems are readily solvable by the application of the dynamic maximization principle. Gould's paper (p. 356) defines a solution to a problem virtually identical to our advertising problem. Given the routine nature of the solution, analytical details are relegated to Appendix A and the dynamics are summarized in the phase diagram labeled Figure 1. Necessary conditions are presented here both for interpretive purposes and for their relevance for the subsequent welfare analysis.

For the advertising problem, necessary conditions for a profit maximum are

$$(7) \qquad G_m = 0$$

$$(8)^9 \qquad -1 + \gamma g'(1-n) \leq 0$$

with equality for $a > 0$

$$(9) \qquad D\gamma = \gamma(\rho + \xi + g(a)) - G$$

$$(10) \qquad Dn = g(a)\,(1-n) - \xi n$$

where $nG \equiv n(pm - cm)$, net production revenues and γ is the current-valued shadow price of n.

In equation (7), the optimal output decision of the firm serves to identify the marginal consumer. We denote optimal monopoly levels of the control variables by an asterisk. Therefore, in equation (7) note that the optimal level of m^* is defined independent of advertising and time. (This latter fact is used in the subsequent welfare analysis.) In long-run equilibrium, the substitution of (9) into (8) yields

$$(11) \qquad g'G(1-n)/(\rho + \xi + g(a)) = 1$$

where the left-hand expression is the discounted marginal value product of disseminative advertising. Not surprisingly, in long-run equilibrium it equals the marginal per capita cost of advertising ($1). Equation (8) yields $a^* = a(\gamma, n)$ with $\partial a^*/\partial\gamma > 0$ and $\partial a^*/\partial n < 0$. An increase in the value of informed consumers promotes advertising; an increase in the size of the informed proportion of the population discourages advertising.

The solution to the demonstration problem has similar characteristics. Necessary conditions for a profit maximum are

$$(12) \qquad nG_m + \gamma\delta(1-n) = 0$$

$$(13) \qquad D\gamma = \gamma(\rho + \xi + \delta m) - G$$

$$(14) \qquad Dn = \delta m\,(1-n) - \xi n$$

Given $\gamma > 0$, $1 > \delta$, $n > 0$, equation (12) implies that output is larger than the static monopoly output, where marginal revenue is set equal to marginal costs (or $nG_m = 0$ in our terms); in this case current marginal costs exceed current marginal revenue by the marginal value product of another informed consumer in terms of the discounted value of future potential net revenue (γ) times the number of additional consumers attracted by the demonstration effect of another customer $[\delta(1-n)]$. This is achieved by lowering price below the level found profitable by the corresponding static monopoly. These lower prices are payments to the current buyers for their advertising of the product. Some comments on the evolution over time of the firm's policy variables in these two problems seem appropriate.

From Figure 1, it is clear that over time, changes in advertising and the proportion of informed consumers are inversely related. Therefore, if the initial proportion of the population informed of the product's existence is small relative to the long-run equilibrium value, then per capita advertising expen-

[9]Provided $\gamma > 0$, $0 < n < 1$, and $g'(0) = +\infty$, $a > 0$ is guaranteed.

ditures should decline through time as the informed proportion of the population grows until the steady state is achieved. This occurs because the number of customers attracted by a given amount of advertising is related to the size of the uninformed segment of the population, which decreases over time.

A slightly different pattern emerges for the demonstration problem, although the analysis is similar. Here, output and price do change. To describe the optimal output (pricing) policy through time, differentiation of equation (12) with respect to time and subsequent substitution from equations (12), (13) and (14) yields

$$DM = G_m[\rho + \xi + \xi/(1-n) - \delta m(1-n)/n + \delta G(1-n)/(nG_m)]/G_{mm} \tag{15}$$

Again, following methods similar to those used in Appendix A, it is possible to draw a phase diagram in (n,m) space.[10] A similar pattern to advertising emerges. Here, if the initial portion of the informed population is low relative to the long-run equilibrium value, then output per informed person (m) is large (price is low) and decreases (price increases) to the long-run steady-state value. This is because the cost of promotion (via lower prices) increases more rapidly as n increases than does the number of customers attracted by this promotion.

II. A Welfare Evaluation of Disseminative Advertising and Demonstration Effects

In both of these problems, we ask a second best policy question: Given the monopoly power of the firm, is the monopoly's information policy socially optimal? This is one of the questions of interest. If the answer is negative, there are implications for market policy. To answer this, we need a measure of social benefits. At any moment of time, per capita social welfare may be measured in a utilitarian fashion as the sum of individual utilities:

$$w = n\int_0^m \alpha(i)\,di + \int_0^1 (Z^i/S)\,di \tag{16}$$

This specification incorporates the assumption (previously stated) that newly informed consumers have the same taste distribution as those already informed. (This means that m is independent of n.) The advertising model is completed by a resource balance equation defined in per capita terms as

$$nmc + \int_0^1 (Z_i/S)\,di + a = r \tag{17}$$

Substituting (17) into (16) allows the per capita net social welfare at each moment of time to be defined in the conventional way:

$$\begin{aligned} w &= n\int_0^m (\alpha(i) - \alpha(m))\,di + nm(\alpha(m) - c) - a + r \\ &\equiv n(I + G) - a + r \end{aligned} \tag{18}$$

where $I \equiv \int_0^m (\alpha(i) - \alpha(m))\,di$

(and G has been previously defined).

From equation (18), net social welfare equals the excess of consumers' utility over that of the marginal consumer (I), plus the monopoly net revenue (G), less advertising resource cost (a), plus the total resources available evaluated at their opportunity cost in terms of $Z(r)$. Thus, our analysis ignores income (or distributional) effects. The only change for the measure of net social welfare for the demonstration problem is that advertising expenditures are deleted from equation (18).

Our procedure for evaluating the social optimality of the monopoly advertising policy is to treat the per capita advertising expenditures parametrically. We measure the sign of the present value of the change in per capita marginal social welfare of advertising, evaluated at the level found to be profitable by the monopoly. This indicates whether increased advertising from the monopoly level raises or lowers social welfare.[11]

For the advertising problem, the profit-maximizing level of m is independent of both advertising and time, so that from equation

[10] Differentiation of equation (15) yields $dm/dn\,(Dm = 0) = [\xi/(1-n)^2 + \delta m/n^2 - \delta G/(n^2 G_m)]/[GG_{mm}/(n G_m^2)] < 0$. This produces a conventional phase diagram where there is a unique optimal trajectory along which m and n move in opposite directions.

[11] The technique is similar to that used by Dixit and Norman.

(18) we may write the per capita marginal social welfare of advertising at each point of time (evaluated at m^* and a^*) as

$$(19) \qquad \left.\frac{dw}{da}\right|_{a^*}^{m^*} = (I^* + G^*)\frac{\partial n}{\partial a} - 1$$

Rearranging terms, we may write the present value of these marginal welfare changes as

$$\Delta W|_{a^*}^{m^*} \equiv \Delta W^I|_{a^*}^{m^*} + \Delta J^a|_{a^*}^{m^*}$$

where

$$(20) \quad \Delta W^I|_{a^*}^{m^*} \equiv \int_0^\infty e^{-(\rho-\eta)t} I^* \frac{\partial n}{\partial a} dt$$

$$(21) \quad \Delta J^a|_{a^*}^{m^*} \equiv \int_0^\infty e^{-(\rho-\eta)t} [G^* \frac{\partial n}{\partial a} - 1]\, dt$$

The term $\Delta W^I(M = m^*, a = a^*)$ is the present value of the marginal consumer benefits, while $\Delta J^a(m = m^*, a = a^*)$ is the present value of the increment to the monopoly's profit. By the definition of the profit-maximizing policy, $\Delta J^a(m = m^*, a = a^*) = 0$.

Following techniques outlined in Hajime Oniki, it is possible to show that for all time periods $\partial n/\partial a > 0$.[12] Therefore, $\Delta W(m = m^*, a = a^*) > 0$, and the monopolist's disseminative advertising falls short of the social optimum when $I^* > 0$. If newly informed consumers were all marginal, then $I^* = 0$ and the monopolist's disseminative advertising would be socially optimal. These results occur because the monopolist is unable to appropriate the inframarginal benefits to the informed consumers. A perfectly discriminating monopolist includes any potential inframarginal benefits in its calculations, collecting these benefits as revenues through its pricing policy. Not surprisingly, therefore, the perfectly discriminating monopolist engages in the socially optimal amount of disseminative advertising.

For demonstration effects, our welfare procedure is similar. The firm's output level, which is time dependent for this problem, is treated parametrically. We measure the sign of the change in the per capita marginal social

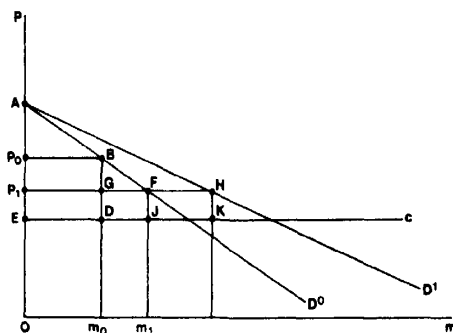



FIGURE 2

D^0 is the initial demand curve; (m_0, P_0) are the initial monopoly output and price; D^1 is the final demand curve with demonstration effects; (m_1, P_1) are the monopoly output and price in the face of demonstration effects; c is the (constant) per unit cost.

welfare of output due to a small increase in output throughout the period, relative to the output level found to be profitable by the monopolist. This measures whether increased information, in the form of lower prices than those of the monopoly, raises or lowers social welfare.

From equation (18), the per capita marginal social welfare of output at each point of time, evaluated at m^*, is

$$(22) \quad \left.\frac{dw}{dm}\right|_{m^*} = n\left(\frac{\partial I^*}{\partial m} + \frac{\partial G^*}{\partial m}\right) + (I^* + G^*)\frac{\partial n}{\partial m}$$

Referring to Figure 2, we see that $n(\partial I^*/\partial m)$, which corresponds to P_0BFP_1, is the gain to informed consumers due to the lower price.

The term $n(\partial G^*/\partial m)$, which corresponds to $(GFJD - P_0BGP_1)$ is the change in current monopoly profits due to the change in price to informed customers.

The term $((I^* + G^*)\ \partial n/\partial m)$, which corresponds to $(AHF + FHKJ)$, are the gains to newly informed consumers and the monopoly due to increased information dissemination.

Rearranging terms, we may write the present value of these marginal welfare changes as

[12]For details see Appendix B.

$$\Delta W|_{m^*} \equiv \Delta W^I|_{m^*} + \Delta J^m|_{m^*}$$

where

(23) $$\Delta W^I|_{m^*} \equiv \int_0^\infty e^{-(\rho-\eta)t} \left(n \frac{\partial I^*}{\partial m} + I^* \frac{\partial n}{\partial m}\right) dt$$

(24) $$\Delta J^m|_{m^*} \equiv \int_0^\infty e^{-(\rho-\eta)t} \left(n \frac{\partial G^*}{\partial m} + G^* \frac{\partial n}{\partial m}\right) dt$$

The term $\Delta W^I(m = m^*)$ is the present value of the marginal consumer benefits, and $\Delta J^m(m = m^*)$ is the present value of the increment to the monopoly's profits. By the definition of the profit-maximizing policy m^*, $\Delta J^m(m = m^*) = 0$. In terms of Figure 2, ΔJ^m is the present value of $(GHDK - P_1BGP_1)$, which at the margin, for the optimal firm output policy, is zero. From the definition of I^*, $n\partial I^*/\partial m = -\alpha' mn > 0$. In long-run equilibrium, from equation (14), $\partial n/\partial m(m=m^*) = \delta(1 - n)/(\delta m^* + \xi) > 0$. Again, following the technique outlined in Oniki, it is possible to show that $\partial n/\partial m > 0$ for all time periods.[13] Therefore, $\Delta W^I(m = m^*) > 0$. In terms of Figure 2, $\Delta W^I(m = m^*)$ is the present value of $(AHF + P_0\,BFP_1)$ which is positive provided $\partial n/\partial m > 0$. As a result, we may conclude that $\Delta W|_{m^*} > 0$. Thus, from the viewpoint of the present value of net social welfare, the monopolist's level of information through reduced prices falls *short* of the social optimum.

The monopolist generally underutilizes demonstration effects as an informational tool for two reasons. First, price reductions are offered to *all* informed customers, not to new (marginal) buyers among the informed. (This corresponds to $n\ \partial I^*/\partial m = -\alpha' mn$.) Thus, the monopolist "pays" existing customers for "word of mouth advertising," which they would undertake anyhow. If the monopolist were able to discriminate by lowering the price only to the marginal customers, this avenue of information diffusion would be used more heavily. This would increase social welfare, as changes in the price to existing customers involve a pure transfer between them and the monopoly with no aggregate welfare effects. Thus, lower prices to the new customers would increase welfare regardless of the price to those who would buy the product in any case.[14] Second, the nondiscriminating monopolist is unable to extract the inframarginal consumer benefits of *newly informed consumers*. (This corresponds to $I^*\ \partial n/\partial m$.) Therefore, he spends too little on this type of promotion even if he were able to discriminate between marginal and inframarginal *informed* consumers. A monopolist that can perfectly discriminate with respect to price should produce socially optimal levels of information through reduced prices and demonstration effects.

III. Advertising as a Substitute for Demonstration Effects

In this section, we combine the models of advertising and demonstration effects discussed in the previous section and ask whether the substitution of disseminative advertising for information dissemination through demonstration effects and lower prices increases or decreases social welfare at the margin.

The procedure follows that of the social evaluation of the magnitude of separate effects in the last section. The equation describing the changes in the informed proportion of the population becomes:[15]

(25) $$Dn = (g(a) + \delta m)(1 - n) - \xi n$$

where $0 < g(a) + \delta m \leq 1$.

The necessary conditions for the profit-maximizing monopoly solution include equations (12) and (25), together with a marginal

[13]The proof is similar to that outlined in Appendix B.

[14]It should be noted that such practices of discrimination between new consumers and existing ones are quite prevalent, although their function may be to induce experimentation rather than to promote diffusion. Periodic "sales," promotional giveaways, tied sales (for example, a razor with blades, a camera with film), and pyramid selling are examples of such practices.

[15]Advertising and demonstration effects are additively separable in this specification of the diffusion process. (In Gould's second model (due initially to Ozga) these effects are interdependent.) Again, this specification is sufficient but not necessary for our welfare results.

advertising equation and a shadow price equation:

$$(26)\quad -1+\gamma g'(1-n)\leq 0,\quad \text{with equality when } a>0$$

$$(27)\quad D\gamma=\gamma(\rho+\xi+g(a)+\delta m)-G$$

The dynamic properties of the solution summarized in Figure 1 remain unaltered by the addition of the advertising to the demonstration effects.

The question we pose is whether or not the allocation of resources to explicit advertising, in the presence of demonstration effects found profitable by a monopolist, is socially excessive or not. To answer this second best question, we evaluate the sign of the marginal social welfare of advertising at that level of advertising and output found profitable by the monopolist, where the monopolist is permitted to adjust his output (price), so as to maximize the discounted stream of profits, given the change in advertising.

From equation (18), the per capita marginal social welfare of advertising at each moment of time evaluated at m^* (the monopolist's profit-maximizing level of output in the presence of advertising) and a^* (the monopolist's profit-maximizing level of advertising) is

$$(28)\quad \left.\frac{dw}{da}\right|_{a^*}^{m^*}=n\left(\frac{\partial I^*}{\partial m^*}\frac{\partial m^*}{\partial a}+\frac{\partial G^*}{\partial m^*}\frac{\partial m^*}{\partial a}\right)+(I^*+G^*)\frac{\partial n}{\partial a}-1$$

where $n((\partial I^*/\partial m^*)(\partial m^*/\partial a)+(\partial G^*/\partial m^*)(\partial m^*/\partial a))$ is the welfare change of all currently informed consumers plus the change in monopoly profits due to a price change, induced by the change in advertising. Moreover $(I^*+G^*)\,\partial n/\partial a$ is the social gain due to the net change in the rate of information dissemination because of an increase in advertising together with induced change in demonstration effects. The last term reflects the marginal cost of a dollar expenditure on advertising.

As before, we define the present value of the change in social welfare as

$$\Delta W\Big|_{a^*}^{m^*}\equiv\Delta W'\Big|_{a^*}^{m^*}+\Delta J^m\Big|_{a^*}^{m^*}$$

where

$$(29)\quad \Delta W'\Big|_{a^*}^{m^*}\equiv\int_0^\infty e^{-(\rho-\eta)t}\left(n\frac{\partial I^*}{\partial m^*}\frac{\partial m^*}{\partial a}+I^*\frac{\partial n}{\partial a}\right)dt$$

$$(30)\quad \Delta J^a\Big|_{a^*}^{m^*}\equiv\int_0^\infty e^{-(\rho-\eta)t}\left(n\frac{\partial G^*}{\partial m^*}\frac{\partial m^*}{\partial a}+G^*\frac{\partial n}{\partial a}-1\right)dt$$

where $\Delta W'$ ($m=m^*$, $a=a^*$) is the present value of the marginal consumer benefits, and ΔJ^a ($m=m^*$, $a=a^*$) is the present value of the increment in monopoly profits.

By definition of the profit-maximizing policies m^* and a^*, ΔJ^a ($m=m^*, a=a^*$) $=0$. From the definition of I^*, $n\,\partial I^*/\partial m^*=-\alpha' n\partial m^*/\partial a$. Over all time periods, $\partial m^*/\partial a<0$. As the advertising partly substitutes for demonstration effects, the monopolist cuts the payment to his customers for their promotional service. In long-run equilibrium, $\partial n^\infty/\partial a>0$ but its sign is ambiguous in other time periods.[16]

This means that the sign of $\Delta W'(m=m^*, a=a^*)$ is ambiguous and no clear-cut welfare statement appears to be possible. As prices rise due to reduced demonstration effects, there are deadweight losses to existing consumers; if advertising (the substitute information technology) expands the pool of informed consumers (and we are certain it does in the long run, but uncertain of the interim and the present value of all these effects), net consumer benefits from newly informed inframarginal consumers may offset this deadweight loss. If all newly informed

[16]By setting $D\gamma=0$, we may substitute for γ in equation (13). This together with equation (28) with $Dn=0$ gives us a two-equation system. Solving yields

$$\partial m^\infty/\partial a=\{g'G_m(g(a)+\xi+\delta m+\rho(1-n))-\delta(1-n)g'G\}/D$$

$$\partial n^\infty/\partial a=-(1-n)\{\delta(1-n)g'G_m-ng'G_{mm}(\rho+g(a)+\xi+\delta m)\}/D$$

where D is positive. This together with the other signs implies that $\partial m^\infty/\partial a<0$ and $\partial n^\infty/\partial a>0$. Following techniques outlined in Oniki and illustrated in Appendix B, it is possible to show that for all time periods $\partial m^*/\partial a<0$ but the sign of $\partial n/\partial a$ is uncertain.

consumers were marginal, then advertising would be excessive.[17]

If we define ϵ_a^I as the elasticity of inframarginal valuations with respect to advertising (permitting price to adjust) and ϵ_a^n as the elasticity of informed consumers with respect to advertising, substitution into equation (31) of these terms tells us that the social optimality of the monopolist's advertising program in present terms rests on the sign of

$$\int_0^\infty e^{-(\rho-\eta)t} \frac{nI^*}{a^*} (\epsilon_{a^*}^I + \epsilon_{a^*}^n)\, dt$$

For example, a monopolist whose optimal advertising policy is such that the sum of these elasticities of opposite sign is zero in each time period would engage in the socially optimal amount of advertising.

IV. Summary and Conclusions

In this paper we analyzed two diffusion processes for the spread of information about the existence, price, and characteristics of a product: advertising and demonstration effects by existing customers. The optimal information policies of a profit-maximizing monopoly were contrasted with those which maximize social welfare.

We demonstrate that for such information spread, accomplished either by explicit advertising campaigns, or through demonstration effects by users of the commodity, monopolists *do not engage* in excess information dissemination from the viewpoint of social resource allocation. In fact, to the extent that new consumers' valuations exceed the marginal valuation, there may be too small an allocation of resources to this activity by a conventional monopolist.

In other words, the socially optimal advertising and pricing policies for these respective effects require a greater speed of information dissemination than that supplied by the monopolist, either in the form of advertising expenditures or reduced prices. Perfectly discriminating monopolists, who capture all the inframarginal benefits through their discriminatory pricing policies, should engage in the socially optimal policies.

When both information technologies are available to the monopolist, the monopolist overutilizes advertising relative to demonstration effects as tools of information dissemination although the absolute amount of information supplied is still too low. Consequently an increase in advertising over and above its profit-maximizing level may or may not be socially desirable. While the substitution of advertising for demonstration effects is socially undesirable, the increased flow of information due to additional advertising is socially beneficial. Thus the introduction of explicit advertising as a substitute for some demonstration effects may or may not be socially excessive as practiced by a monopolist. The final judgement rests on a tradeoff between the deadweight losses to inframarginal consumers and the potential gains to newly informed consumers through increased information dissemination.

These conclusions are valid only for advertising which supplies *correct* information. While no one would disagree that some advertising supplies such information, it is not generally agreed that *all* advertising plays this role. The validity of our analysis of advertising must be restricted to situations where consumers are entirely ignorant of the product or where the quality aspects of a commodity are readily and inexpensively verifiable prior to purchase, or where quality is well defined and quickly and cheaply measured with use. To fully evaluate the social utility of advertising it is necessary to investigate its role under conditions where verification of advertised qualities is difficult and/or expensive, so that a possibility for misleading advertising exists.

Appendix A: Details of the Solution to the Informative Advertising Problem

To maximize the discounted stream of profits, subject to changes in the proportion of the informed population, form the current value Hamiltonian:

[17] If newly informed consumers have lower preference for the good in question than those of previously informed consumers, I^* is likely to be small and therefore advertising is more likely to be excessive than otherwise. (See fn. 5 above.)

$$H = nG(m) - a + \gamma[g(a)(1-n) - \xi n]$$

Necessary (and sufficient given concavity) conditions for a maximum of H are

$$(A1)\quad G_m = 0$$

$$(A2)\quad -1 + \gamma g'(1-n) = 0$$

$$(A3)\quad D\gamma = \gamma(\rho + \xi + g(a)) - G$$

$$(A4)\quad Dn = g(a)(1-n) - \xi n$$

Equation (A1) yields m^*, independent of advertising and time.

Taking the time derivative of (A2) and substituting (A3) and (A4) yields

$$(A5)\quad Da = g'(-(\rho + \xi/(1-n)) + Gg'(1-n))/g''$$

For the phase diagram of Figure 1, consider (A5) and (A4) evaluated at m^* and a^* at rest. Then,

$$\left.\frac{da}{dn}\right|_{Da=0} = \frac{(\xi/(1-n) + Gg')}{Gg''(1-n)} < 0$$

$$\left.\frac{da}{dn}\right|_{Dn=0} = \frac{g+\xi}{g'(1-n)} > 0$$

Therefore, $Da = 0$ is negatively sloped while $Dn = 0$ is postively sloped. Further, the optimal trajectory indicated in Figure 2 has the conventional stability properties given the assumptions of the model.

Appendix B: Illustration of the Comparative Dynamics—The Effect of Changing Levels of Advertising

Consider equations (7), (9), and (10). Advertising is treated parametrically so that equation (8) is irrelevant at this point. From equation (7), m^* is independent of advertising and therefore does not change. In long-run equilibrium from equations (9) and (10), we may show that:

$$(A6)\quad \left.\frac{\partial\gamma^\infty}{\partial a}\right|_{D\gamma=0} = -\frac{\gamma^\infty g'}{(\rho+\xi+g)} < 0$$

$$(A7)\quad \left.\frac{\partial\nu^\infty}{\partial a}\right|_{Dn=0} = \frac{(1-n)g'}{(\xi+g)} > 0$$

For other time periods, differentiation of equations (9) and (10) yields:

$$(A8)\quad \begin{bmatrix} \frac{\partial\gamma}{\partial a} \\ \frac{\partial n}{\partial a} \end{bmatrix} = \begin{bmatrix} (\rho+\xi+g) & 0 \\ 0 & -(\xi+g) \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial\gamma}{\partial a} \\ \frac{\partial n}{\partial a} \end{bmatrix} + \begin{bmatrix} \gamma g' \\ (1-n)g' \end{bmatrix}$$

An examination of the feasible sign configurations in (A8), together with (A6) and (A7) and $\partial n^0/\partial a = 0$ reveals that $\partial n/\partial a > 0$ for all time periods. (The sign of $\partial\gamma/\partial a$ is irrelevant for the analysis.)

References

I. Bernhardt and K. D. Mackenzie, "Some Problems in Using Diffusion Models for New Products," *Manage. Sci.*, Oct. 1972, *19*, 187–200.

J. S. Coleman, E. Katz, and H. Menzel, *Medical Innovation: A Diffusion Study*, New York 1966.

A. Dixit and V. Norman, "Advertising and Welfare," *Bell J. Econ.*, Spring 1976, *9*, 1–17.

James Ferguson, *Advertising and Competition: Theory, Measurement, Fact*, Cambridge, Mass. 1974.

John Kenneth Galbraith, *The New Industrial State*, Boston 1967.

S. Glaister, "Advertising Policy and Returns to Scale in Markets Where Information is Passed Between Individuals," *Economica*, May 1974, *41*, 139–156.

J. P. Gould, "Diffusion Processes and Optimal Advertising Policy," in Edmund S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970, 338–68.

N. Kaldor, "The Economic Aspects of Advertising," *Rev. Econ. Stud.*, No. 1, 1950, *18*, 1–27.

E. Katz, M. L. Levin, and H. Hamilton, "Traditions of Research on the Diffusion of Innovation," *Amer. Soc. Rev.*, Apr. 1963, *28*, 237–52.

Y. Kotowitz and F. Mathewson, "Advertising,

Consumer Information and Product Quality," work. paper no. 7801, Instit. Pol. Analysis, Univ. Toronto.

P. Nelson, "Information and Consumer Behavior," *J. Polit. Econ.*, Mar./Apr. 1970, *78*, 311–29.

——— "Advertising as Information," *J. Polit. Econ.*, July/Aug. 1974, *82*, 729–54.

H. Oniki, "Comparative Dynamics in Optimal Control Theory," *J. Econ. Theory*, June 1973, *6*, 265–83.

S. Ozga, "Imperfect Markets Through Lack of Knowledge," *Quart. J. Econ.*, Feb. 1960, *74*, 29–52.

Richard Schmalensee, *The Economics of Advertising*, Amsterdam 1972.

G. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–75; reprinted in his *The Organization of Industry*, Homewood 1961, 171–90.

L. Telser, "Advertising and Competition," *J. Polit. Econ.*, Dec. 1964, *72*, 537–62.

Measurement Errors and the Permanent Income Hypothesis: Evidence From Rural India

By SURJIT S. BHALLA*

The traditional, or Keynesian, consumption hypothesis postulated that (a) current income was a prime determinant of consumption; and (b) that the average propensity to consume declined with income. This hypothesis was rejected by Milton Friedman's permanent income hypothesis (*PIH*) which contended that (a) permanent income, and not current income, was the relevant determinant of consumption; and (b) that permanent consumption was proportional to permanent income (the proportionality hypothesis).[1]

This paper attempts to test (and distinguish between) the two theories. Answers to two questions are sought: How important is the current income-permanent income distinction; and is the proportionality hypothesis, an important and controversial aspect of Friedman's theory, valid? The resolution of the former question has important implications for the understanding of habit and lags in consumption behavior and the efficacy of short-run macro-economic policies. The validity of the proportionality hypothesis has an important bearing on the recently popular controversy regarding the tradeoff between growth and equity in developing countries.[2]

The above two questions have been extensively investigated since the publication of Friedman's theory two decades ago.[3] Nevertheless, with one exception, *all* these tests (including Friedman's) have been indirect and nonrigorous. The exception was an ingenious study by Nissan Liviatan, who used two-year panel Israeli data to test the *PIH*, and, in particular, the proportionality hypothesis.[4] His study, however, suffered from a major drawback in that it failed to account for the bias caused by common measurement errors in consumption and income. This paper extends Liviatan's analysis by explicitly allowing for the presence of measurement errors in all the variables—income, consumption, and saving. The empirical estimation of the variances of these errors (source of bias) is possible due to the availability of *independent* estimates of consumption, savings, and income.[5] The data base is a three-year panel survey, conducted by the National Council of Applied Economic Research (*NCAER*), New Delhi, of 4,118 households in rural India. (See Appendix A for description of data and definitions of variables.)

The incorporation of error variances in the

---

*World Bank. This is a shortened version of a paper which was written while I was a recipient of a Rockefeller Post-Doctoral Fellowship at the Rand Corporation. The financial support is gratefully acknowledged. I would like to thank William Rogers for helpful discussions, Thomas Mayer and Charles E. Phelps for comments on an earlier draft, and David Harrison for research assistance. The views expressed in this paper are mine alone

[1]Analogous conclusions are reached by the Modigliani-Brumberg life cycle hypothesis (*LCH*) of consumption behavior. The two theories, *LCH* and *PIH*, use different empirical methods but employ a common theoretical model of consumer behavior. The empirical techniques used in this paper apply directly to Friedman's formulation of the theory. Hence, the emphasis is on testing the *PIH*.

[2]If, as the proportionality hypothesis implies, the propensity to consume permanent income is independent of the level of permanent income, then redistribution policies will be neutral with respect to savings, and therefore growth. (This is predicated on the assumption that growth in less developed countries (*LDCs*) is constrained by a lack of finance rather than of investment opportunities.) Thus, the *PIH* contends that there is *no* tradeoff between redistribution (equity) and growth.

[3]An excellent summary and interpretation of the voluminous literature on consumption behavior is contained in Thomas Mayer's book.

[4]Friedman accepted the compelling logic of Liviatan's tests and stated that "... if these [Liviatan's] results should be confirmed for other bodies of data, they would constitute relevant and significant evidence that the elasticity of permanent components is less than unity [the proportionality hypothesis]" (1963a, p. 63).

[5]These independent estimates allow for *two* measures of consumption—direct consumption and residual consumption (income minus savings). This "excess" of information is what makes the estimation of error variances possible. See Section II (and Appendix B) for details of methodology.

analysis, a unique aspect of this study, insures that unlike Liviatan's tests, the tests of the *PIH* in this paper are theoretically correct and most rigorous to date. Further, accounting for errors allows for the estimation of unbiased parameters of the Keynesian consumption function and makes possible a valid comparison with the *PIH*.

The plan of the paper is as follows. Section I presents a version of the *PIH* which is generalized to include the presence of pure measurement errors. Estimates of these error variances, as well as the uncorrected and corrected parameters of a Keynesian consumption function, are also presented. The "horizon," a crucial parameter in the *PIH*, is defined, estimated, and corrected for measurement errors in Section II. A test of the proportionality hypothesis, and a comparison of the *PIH* with the standard Keynesian model, is presented in Section III.

I. A Generalized Version of the *PIH*

An important feature of this study is the distinction it makes between measurement errors proper, and the "measurement error" caused by the presence of transitory terms in the measured variables. Accounting for this distinction is necessary, and crucial, for a valid test of the *PIH*.

If measurement errors in income y, consumption c, and savings s are present, then a generalized version of the permanent income hypothesis would be

$$\text{(1a)} \quad z^* = z' + z''$$

$$\text{(1b)} \quad z = z^* + z^0, \qquad z = y, c, s$$

where z^* represents the true value, $z'(z'')$ represent the permanent (transitory) terms, z represents the measured value and z^0 the measurement error. Analogously, the variables can be expressed in logarithmic terms:[6]

$$\text{(2a)} \quad Z^* = Z' + Z''$$

$$\text{(2b)} \quad Z = Z^* + Z^0 \qquad Z = Y, C$$

The *PIH* assumes that the transitory components are independent of the permanent values, and each other, i.e.,

$$\text{(3)} \quad cov\,(C', C'') = cov\,(Y', Y'') = cov\,(C'', Y'') = 0$$

Analogous to (3), it may be assumed that measurement errors per se are not correlated with the true values, or with each other:[7]

$$\text{(4)} \quad cov\,(C^*, C^0) = cov\,(Y^*, Y^0) = cov\,(C^0, Y^0) = 0$$

Apart from equations (2) and (3), a third relationship is postulated by the *PIH*; namely, that the permanent components are systematically related to each other, $c' = ky'$, where k is assumed to be dependent on interest rates, tastes, compositon of wealth, etc., but *independent* of y'. This assertion implies a unitary elasticity between the permanent components, i.e.,

$$\text{(5)} \quad C' = K + Y'$$

As is well known, estimation of the standard Keynesian function ($c = a + by$ or $C = A + BY$) in the absence of measurement errors yields a downwardly biased estimates of the permanent elasticity, η':

$$\text{(6)} \quad \eta = var\ Y' \,/\, (var\ Y' + var\ Y'') < \eta' = 1$$

Allowance for measurement errors in Y biases

[6] Lowercase letters represent natural numbers: upper case letters represent the logarithm (*log*) of a number. Equation (1) (arithmetic model) is implied by the assumption that transitory terms are additively distributed; if the variables are multiplicatively distributed (i.e., $y = y'y''$) then a logarithmic model results. The former assumption results in a point estimate of the marginal propensity to consume; the latter in a point estimate of the consumption elasticity. Though both models are discussed, the development of the model is outlined in its logarithmic variant. Note that equation (2) does not contain an equation for savings, which can be negative for an individual household

[7] The assumption that the measurement errors are not correlated with each other needs to be qualified. Some items are common to consumption and income (for example, housing) whereas others are common to savings and income (nonmonetized investment). The magnitudes of these covariances are likely to be relatively small and are therefore ignored.

downward the true measured elasticity η^*, as well:

$$(6') \quad \eta^* > \eta = cov\,(C, Y) \,/\, (var\,Y^* + var\,Y^0)$$

The relative magnitudes of $var\ Y''$ and $var\ Y^0$ are unknown; lack of knowledge of the error variance $var\ Y^0$ can negate any comparison between the two consumption theories; that is, a comparison of the permanent elasticity η' with the traditional elasticity η^*. Further, a valid test of whether η' is equal to unity cannot be conducted. Fortunately, given certain assumptions, the magnitude of $var\ Y^0$ (and error variances in savings and consumption) can be estimated.

A maintained and general assumption in this paper is that measurement errors (c^0, s^0, and y^0) are independent of the true values, and have a zero mean. Thus, $E(z) = E(z^*)$ and $E(c) = E(y) - E(s)$, where $E(\ )$ is the expectation operator. Thus, the aggregate estimates of c and its residual estimate, c_r, $c_r = y - s$ (alternatively s and s_r), should be approximately equal. Such is not the case. The saving rates for the three years—1968–69, 1969–70, and 1970–71—are −0.6, 4.2, and 5.4 percent, respectively, for the direct measure, and 13.6, 14.6, and 16.3 percent, respectively, for the residual measure s_r. ($s_r = y - c$ where c is (presumably) understated.) The difference between s and s_r is too large to be accounted for by the omission of certain savings (cash and jewelry) from s. The explanation is more likely in terms of the particular aims of the survey—detailed assessment of investment patterns (and therefore, s) was one of the major goals of the survey; enumeration of consumption expenditures (and therefore s_r) was not. Thus, it is likely that a certain fraction of consumption expenditures was systematically excluded from enumeration and therefore $s_r > s$. Comparison with published national estimates also indicates that s at the aggregate level is a more accurate measure of rural savings.[8]

An equality between the two estimates of savings can be achieved *ex post* by assuming that each household underestimates its consumption by a constant proportion α. Thus, direct consumption c is no longer equal to $c^* + c^0$, rather, $c = \alpha c^* + c^0$, where $0 < \alpha < 1$. If s is a "true" measure of aggregate savings, then $\alpha = E(c)/[E(y) - E(s)]$.

Given the *ex post* equality between c and c_r, and the value of α, a "method of moments" approach can be used to derive the variances, $var\ y^0$, $var\ c^0$, and $var\ s^0$. (A key equation in the derivation is the identity between true values $y^* = c^* + s^*$.) Two different assumptions about the distribution of the errors are considered: (a) errors affect true values additively (arithmetic model, equation (1b)); and (b) errors affect true values proportionally (multiplicative model, equation (2b)). These two assumptions, though not exhaustive, cover a wide range of possibilities. Of the two, it is likely that the multiplicative assumption is more reasonable, that is, it is more likely that rich and poor households make the same *proportional* error, X percent, rather than the same *absolute* error.

The solution of error variances is straightforward for the arithmetic model; for the more realistic multiplicative model, the methodology is much more involved. The difficulty arises primarily because the crucial identity $y^* = c^* + s^*$ does not exist in its (multiplicative) logarithmic transformation, i.e., $Y^* \neq C^* + S^*$. Nevertheless, an appropriate procedure exists. The solutions are outlined in Appendix B. The expressions for the income error variances for the two models are[9]

$$(7a) \quad var\,y^0 = var\,y - cov\,(s, y) - \alpha^{-1}\,cov\,(c, y)$$

$$(7b) \quad var\,Y^0 = ln\,(1 + var\,M)$$

where $var\ M =$

$$\frac{\alpha\,var\,y - cov\,(c, y) - \alpha\,cov\,(s, y)}{\alpha\,[E(y)]^2 + cov\,(c, y) + \alpha\,cov\,(s, y)}$$

[8]See the author for a detailed discussion of various savings estimates, and a comparison of s and s_r.

[9]Appendix B also contains expressions for errors in savings and consumption. Since the primary concern is with errors in the independent variable, income, only these are discussed in the rest of the paper.

TABLE 1—ERROR VARIANCES AND RELATED STATISTICS

	Year 1 1968–69	Year 2 1969–70	Year 3 1970–71
Mean Income, Rs.	4544	4732	4988
Alpha $E(c)/[E(y) - E(s)]$	.804	.850	.860
Direct Consumption, c			
APC	.778	.759	.741
MPC, β	.39	.49	.47
β (adjusted by α)	.48	.58	.55
η (arithmetic model)	.50	.65	.63
η (*log* model)	53	.71	.73
Residual Consumption, c_r			
APC	.97	.89	.86
β	91	.70	.65
η (arithmetic model)	.94	.79	.76
η (*log* model)	88	.80	.80
Errors and True (corrected) Values			
Error Ratio ($var\ y^0/var\ y$)	43.2	12.4	10.4
Typical Error (Rs.)	3014	1429	1430
Error Ratio ($var\ Y^0/var\ Y$)	26.3	8.6	8.0
Typical Error (Percent)	49.6	23 2	21.7
True *MPC*, β^*	.84	.66	.61
True Elasticity, η^* (arithmetic model)	.87	.74	.71
True Elasticity, η^* (*log* model)	.72	.78	.80

Source: *NCAER* survey, cultivators only For method of selection of observations, see Appendix A.

Equation (7) allows estimation of the error variances in terms of observed quantities; the estimates of these errors for the three years, and related data, are reported in Table 1. The results pertaining to the source of bias, the income error ratio, are striking. Errors (arithmetic model) account for as much as 43 percent of income variance in the first year of the survey; these decline significantly to 12 and 10 percent in the following two years. Even the more realistic assumption of proportional errors yields large errors for the first year—26 percent. Again, the error variance drops radically—9 and 8 percent in the second and third year, respectively.

A direct interpretation of measurement errors can be in terms of "typical" errors. Defining this to be one standard deviation, the errors are uncomfortably large: Rs. 3,014, Rs. 1,429, and Rs. 1,430 for the arithmetic model, and 50, 23, and 22 percent for the *log* model. Given that mean income is approximately Rs. 4,500, the error for the first year is especially discomforting.

The large error variance for the first year could have been caused by respondents being unfamiliar with survey procedures, suspicious of interviewers, etc. If these errors are typical of one-shot surverys in *LDC*s, then the analysis based on them is suspect.

There is no basis for judging these results since *no* prior estimates exist. The importance of the magnitude of even the low error variance for the third year, 8–10 percent, is emphasized by the fact that Friedman (1963a) contended that an error of 6 percent was sufficient to at least cast doubt upon, if not reverse, some of the results reached by Liviatan.

The measurement error variances are used to correct the observed coefficients and the results are also reported in Table 1. For the year 1970–71, the true marginal propensity to consume (*MPC*) β^* is 0.61 and the true elasticity (*log* model) η^* is .79. The observed *MPC* ranges between .55 and .65 (direct and residual estimates), and the observed elasticity between .73 and .80. The result pertaining to η^* is in close correspondence with other data and the Keynesian hypothesis, that is, the average propensity to consume (*APC*) of households declines with increases in current

income. If the *PIH* is mostly true, then the permanent elasticity should be significantly higher than η^* and (perhaps) equal to unity. But before η' can be computed, it is necessary for the horizon to be estimated. This crucial parameter can also be biased by measurement errors; uncorrected and corrected estimates of the horizon are presented in the next section.

II. The Horizon

The proper length of the horizon has been a contentious issue in the literature on consumption theory.[10] Part of the disagreement can be attributed to the fact that Friedman offers *two* definitions of the horizon without attempting to show any equivalence between the two. One definition is derived from considerations of wealth and its conversion into permanent income. Accordingly, if the subjective discount rate is r, "the horizon can then be defined as $1/r$, or 'the number of years purchase' implied by the discount rate" (Friedman, 1963b, p.7). The second definition is directly related to the statistical formulation of the *PIH* and concerns itself with the separation of an income stream into its permanent and transitory components; "the length of time a factor must affect income before it is regarded as permanent is a measure of the length of the horizon" (Holbrook, p.750). According to the latter definition, a two-year horizon would imply that the transitory terms for the two years are not correlated. Analogously, "suppose the horizon were three years. Some factors would then be regarded as transitory even though they affected income in two years, so transitory terms in successive years would be correlated, though in years separated by a year they would be uncorrelated" (Friedman, 1957, p. 196).

A direct estimate of the second definition of the horizon can be achieved with *NCAER* data. Assumptions (additional to equations (1)–(5)) are necessary, but since these are identical to those made by Friedman, the tests are within the spirit (if not the letter) of the *PIH*. In particular, it is assumed that the permanent component of income changes (if at all) in a systematic manner from year to year. The systematic nature of change is either "that permanent components maintain the same ratio to the mean of the group in different years" (mean assumption), or "that the fraction of the total variability contributed by the permanent components is the same in successive years" (variability assumption), (Friedman, 1957, p. 184).

For the *log* model, *mean assumption*, the elasticity of incomes for any two years of the survey (indicated by the subscripts i and j), is

$$\begin{aligned}(8)\quad \eta_{Y_iY_j} &= cov\,(Y_i, Y_j)/var\,Y_j \\ &= cov\,(Y_i' + Y_i'' + Y_i^0,\ Y_j' + Y_j'' + Y_j^0)/var\,(Y_j' + Y_j'' + Y_j^0)\end{aligned}$$

If Y_i' is systematically related to Y_j' (for example, $y_i' = \lambda y_j'$ or $Y_i' = \Gamma + Y_j'$), then

$$\begin{aligned}(8')\quad \eta_{Y_iY_j} &= cov\,(\Gamma + Y_j' + Y_i'' + Y_i^0,\ Y_j' + Y_j'' + Y_j^0)/var\,Y_j \\ &= (var\,Y_j' + \Delta Y_i''Y_j'' + \Delta Y_i^0 Y_j^0)/var\,Y_j\end{aligned}$$

where Δ's represent the respective covariances (i.e., $\Delta Y_i''Y_j'' = cov\,(Y_i'', Y_j'')$). If these covariances are assumed to be zero, then (8′) reduces to

$$(9)\qquad \eta_{Y_iY_j} = var\,Y_j'/var\,Y_j$$

Equation (9) is exactly as that obtained for the elasticity of consumption with respect to measured income, η (equation (6)). Analogously, it can be shown that the correlation of incomes $\rho_{Y_iY_j}$ is the measure comparable to η under the *variability assumption* (see Friedman, 1957, p. 184):

$$(9')\quad \rho_{Y_iY_j} = cov(Y_i, Y_j)/\{(var\,Y_i)^{.5}(var\,Y_j)^{.5}\}$$

This implies $\eta = cov(C_j, Y_j)/var\,Y_j$. These "equivalences" can now be used to test for the length of the horizon. If transitory incomes

[10] Friedman contends that the horizon is three years. Though Khan Mohabatt and Evangelos Simos support Friedman, Robert Holbrook asserts that the horizon is considerably shorter. Michael Landsberger, however, finds evidence in support of a three-year horizon with Israeli data.

TABLE 2—EVIDENCE ON CONSUMER HORIZON—CULTIVATOR HOUSEHOLDS, RURAL INDIA

	Elasticities			Horizon	Correlations		Horizon
	η	$\eta_{y_2y_3}$	$\eta_{y_1y_3}$	Mean Assumption	$\rho_{y_2y_3}$	$\rho_{y_1y_3}$	Variability Assumption
Arithmetic Model							
Direct Consumption	.63	.70	.69	$H > 3$	.72	.61	$2 < H < 3$
Residual Consumption	.76	.70	.69	$H < 2$	.72	61	$H < 2$
True Consumption	.71	.89	.57	$2 < H < 3$	.93	.43	$2 < H < 3$
Log Model							
Direct Consumption	.73	.77	.83	[a]	.75	.66	$2 < H < 3$
Residual Consumption	.80	.77	83	[a]	.75	.66	$H < 2$
True Consumption	.794	[b]	[b]	[b]	.82	.799	$H = 3$

[a]Coefficients move opposite to that "predicted." See fn. 12

[b]Correction of $\eta_{y_iy_j}$ is complex for *log* values if intertemporal correlation in measurement errors, equation (11), is allowed. The solution is not attempted.

are correlated, then for incomes measured *within* the horizon $\eta_{Y_iY_j} \neq \eta$ (a positive correlation implies $\eta_{Y_iY_j} > \eta$). Progressively larger distances between i and j should decrease the correlation towards zero. At the endpoints of the horizon, the correlation between transitory terms is equal to zero by definition, and $\eta_{Y_iY_j}$ (or $\rho_{Y_iY_j}$) equals η.

The discussion above is for transitory incomes proper. The presence of measurement errors can severely bias the above methodology and yield erroneous estimates of the horizon H. Equations (8′) and (9′) indicate that there are two error variances affecting the comparison: the measurement errors present in incomes for years i and j; and the intertemporal covariance between the errors, $cov(Y_i^0, Y_j^0)$. Estimates for individual year errors $var\ Y_i^0$ and $var\ Y_j^0$ are available from Table 1. The intertemporal covariance $cov(Y_i^0, Y_j^0)$ may be considered insignificant since it involves correlations in errors a year, and two years, apart. However, it is possible that some errors are repeated and thus $cov(Y_i^0, Y_j^0) \neq 0$. An estimate of this covariance *is* possible if it is assumed that the intertemporal correlation of measurement errors is nonzero for the *same* variable and zero otherwise, i.e.,

$$(10)\quad cov(w_i^0, z_j^0) \neq 0 \quad w = z; \qquad = 0 \quad w \neq z \qquad w, z = c, s, y; \quad i, j = 1,2,3$$

Incorporation of assumption (10) yields the following estimate for $cov(y_i^1, y_j^0)$:[11]

$$(11)\quad cov(y_i^0, y_j^0) = cov(y_i, y_j) - cov(s_i, y_j) - \alpha^{-1}\, cov(c_i, y_j)$$

The measurement error corrected estimates of H are presented with the uncorrected estimates in Table 2. These estimates are derived for the mean assumption (comparison of η with $\eta_{y_iy_j}$) and variability assumption (comparison of η with $\rho_{y_iy_j}$).[12] It is seen that the "traditional," uncorrected (and biased) direct consumption estimate yields the result that the horizon (arithmetic model, mean assumption) is greater than three years. This result is in direct contradiction with most other estimates of the horizon. The importance of accounting for measurement errors is

[11]Derivation of equation (11) is contained in the author's working paper. The analogous solution for the *log* model, i.e., $cov(Y_i^0, Y_j^0)$ is more complex and is not attempted. The complexity arises from the fact that savings can be zero or negative, and as such cannot be (*log*) transformed.

[12]No estimates for the *log* model, mean assumption, are presented, since the elasticity two years apart $\eta_{Y_1Y_3}$ is *greater* than the elasticity one year apart $\eta_{Y_2Y_3}$. This "opposite" movement is easily explained if measurement errors in income are negatively correlated intertemporally. As mentioned in fn. 11, a test of the sign and magnitude of this covariance is complex and not attempted in this paper. Since estimates for the arithmetic model are available, this "error of omission" does not appear to be restrictive.

emphasized by the fact that correction for these errors reduces the estimate of H to between two and three years. The variability assumption, *log* model, yields an estimate of H equal to three years. All other estimates (corrected and uncorrected) place H to be less than three years.

The "corrected" result that the upper bound for the horizon is three years is used in the next section to conduct strong tests of two components of the *PIH*: (a) how much larger is the permanent elasticity η' from the corrected elasticity η^* (i.e., how important is the distinction between permanent and transitory income), and (b) is the proportionality hypothesis valid, i.e., is $\eta' = 1$. The tests of the horizon itself were based on the assumption that η' *is* equal to unity. However, a "logical trap" is avoided if one notes that the method of estimating H—comparison of $\eta_{Y_i Y_j}$ and η—yields an upper bound of the horizon if the permanent consumption-income elasticity is less than unity. Since the bias is in the "right" direction, the rigorous tests outlined next are valid.

III. Estimates of the Permanent Income Model

A generalized expression for the measured consumption elasticity, with *PIH* assumptions, is

$$\eta = var\ Y'/(var\ Y' + var\ Y'' + var\ Y^0) < \eta^* < \eta' = 1$$

The fact that it is the presence of errors (Y'' and Y^0) in Y which causes η' to be downwardly biased suggests the use of an instrumental variable for Y. If this variable is represented by X, then the (instrument) consumption elasticity is

$$\eta^X = cov(C, X)/cov(Y, X)$$

If the proportionality hypothesis is valid ($C' = K + Y'$), then

$$(12)\quad \eta^X = (cov(Y', X') + \Delta C''X'' + \Delta C^0X^0)/(cov(Y', X') + \Delta Y''C'' + \Delta Y^0X^0)$$

where the Δ's represent the covariances between the transitory terms and measurement errors. And if the Δ's are zero, then η^X reduces to the unitary permanent elasticity estimate η'.

Liviatan used the instrumental variable technique with two-year Israeli data and concluded that η^X was *less* than unity. His instruments were the estimates of consumption and income of the "other" year. Unfortunately, his estimate of consumption was a residual estimate (i.e., $c_r = y - s$) and therefore common measurement errors were present in c and y. Further, two-year data meant (if $H > 2$) that the Δ's were nonzero. Consequently, Liviatan's results were questioned by Friedman. Nevertheless, Liviatan's tests were accepted by Friedman, who stated that "it is most desirable that his [Liviatan's] analysis be applied to data not marred by common errors of measurement in income and consumption. It would further be highly desirable to have such data spanning at least a three-year period so that income and consumption for one year could be used as an instrumental variable for a year at least two years later or earlier" (1973b, p. 63).

Given the result that the upper bound for the horizon is three years, the three-year *NCAER* panel data fulfills all the requirements for a valid test. Further, the availability of separate estimates for consumption and saving suggests the presence of six proper instruments: income and two consumptions definitions for year 1 (c and c_r) for *each* definition of consumption for year 3. If the *PIH* is correct, each instrument should yield an unbiased unitary permanent elasticity. This is only possible if all the Δ's are zero. The correlation amongst transitory terms is zero (by definition) for variables two years apart.[13] The intertemporal measurement error correlations, however, may be nonzero. If so, then the use of first-year consumption C_1 (subscripts denote years) on a regression involving

[13]The results of Table 3 are for a definition of consumption which *excludes* durable expenditures. This exclusion should minimize any possible correlation between the transitory components. All the tests of Table 3 were reestimated for a definition of consumption which included durable expenditure; virtually no difference was observed in the results.

TABLE 3—CONSUMPTION FUNCTION PARAMETERS: ALL HOUSEHOLDS, 1970–71 ($N=2453$)

	Arithmetic Model				Logarithmic Model	
	Residual		Direct		Residual	Direct
	β	η	β	η	η	η
Measured	.65	.76	.55	.63	.80	.73
	(.007)	(.010)	(.008)	(.010)	(.008)	(.008)
Corrected (β^*, η^*)	.61	.71	.61	.71	.79	.79
Permanent	.77†	.89†	.73	.85	.89†	.86
(Instrument: c_1)	(.015)	(.020)	(.017)	(.020)	(.012)	(.013)
Permanent	.70	.81	.66	.76	.85	.80
(Instrument: c_{r_1})	(.013)	(.018)	(.015)	(.018)	(.012)	(.013)
Permanent	.71	.82	.65	.76	.84	.79
(Instrument: y_1)	(.012)	(.016)	(.014)	(.016)	(.012)	(.012)

Note: † refers to the "preferred" instrumental variable estimate See text for details. Figures in parentheses are the standard errors of the coefficients. Arithmetic model. reported means and standard errors for β have been adjusted by α; reported elasticities η are computed at the mean value for consumption. The standard errors are valid under the assumption that the *APC* is known, and is equal to a constant, k

C_3 and Y_3, for instance, would imply that η^X would be biased due to the presence of $\Delta C_1^0 C_3^0$; similarly, if Y is used as an instrument, i.e., $\Delta Y_1^0 Y_3^0 \neq 0$. However, if the assumption of equation (10) is accepted, then an instrument does exist whose use would yield an unbiased estimate of the permanent elasticity, η' (i.e., Δ's $= 0$). This instrument is direct consumption, year 1, on a regression between residual consumption and income, year 3, i.e., $\eta\dagger$ where

$$\eta\dagger = cov(C_{r_1}, C_1)/cov(Y_1, C_1)$$

In this instance, only nonsimilar variables are involved and the correlation amongst these errors is assumed to be zero. Thus,

$$\begin{aligned}\eta\dagger &= (cov(Y_3', C_1') + \Delta C_3'' C_1'' + \Delta C_3^0 C_1^0)/\\ &\quad (cov\,(Y_3', C_1') + \Delta Y_3'' C_1'' + \Delta Y_3^0 C_1^0)\\ &= cov\,(Y_3', C_1')/cov\,(Y_3', C_1') = 1\end{aligned}$$

Table 3 presents the results for both the arithmetic and *log* model of consumer behavior, and for all six instruments for income year 3. The two models yield point estimates of the permanent marginal propensity to consume, β', and the permanent consumption elasticity, η'.[14]

Estimates of the true measured propensity and elasticity (from Table 1) are also presented. The results of Table 3 strongly reject the proportionality hypothesis. *All* equations, for both models, yield an elasticity which is considerably less than unity. However, the *PIH* estimates do tend to be larger than the corresponding Keynesian, current income estimates. The preferred estimate $\eta\dagger$ is 25 percent larger, arithmetic model, and 13 percent larger, *log* model, than the corresponding traditional or Keynesian elasticity, η^*. Thus, the general conclusion that emerges is that Friedman's contention of a "lag" in consumption behavior is justified, but *not* the assertion that permanent consumption is proportional to permanent income.[15]

An aggregative relationship as estimated in Table 3 may not be valid. This is due to the presumed presence of "subsistence" households in rural India. These households are definitionally constrained to consume their entire income, i.e., their $APC = k = 1$.[16] Households beyond the subsistence level presumably have an $APC < 1$. An explicit

[14]The estimation of η' for the arithmetic model is computed under the assumption that the average propensity to consume k is a known constant. Thus, $\eta' = \beta^x/k = \beta'/k$.

[15]All tests reported in Tables 3 and 4 were conducted in reverse, i.e., third-year variables as instruments for the first-year relationship. The results were qualitatively identical to those reported in this paper. These results further strengthen the conclusions of this paper.

[16]Arnold Zellner discusses this "weakness" or nongenerality of the permanent income hypothesis.

TABLE 4—CONSUMPTION FUNCTION PARAMETERS

	Arithmetic Model				Logarithmic Model	
	Residual		Direct		Residual	Direct
	β	η	β	η	η	η
Subsistence Households (N = 1146)						
Measured	.86	.85	.84	.83	.83	.80
	(.012)	(.012)	(.014)	(.013)	(.013)	(.014)
Corrected (β^*, η^*)	.86	.86	.86	.86	.82	.82
Permanent	1.01†	.997†	1.03	1.03	1.03†	1.04
(Instrument: c_1)	(.026)	(.026)	(.030)	(.027)	(.029)	(.040)
Permanent	.97	.96	.97	.96	.96	.95
(Instrument: c_{r_1})	(.025)	(.025)	(.028)	(.026)	(.029)	(.03)
Permanent	.99	.98	.99	.98	.97	.94
(Instrument: y_1)	(.024)	(.024)	(.027)	(.025)	(.030)	(.031)
Nonsubsistence Households (N = 1307)						
Measured	.62	.76	.50	.62	.80	.72
	(.011)	(.014)	(.013)	(.013)	(.013)	(.014)
Corrected (β^*, η^*)	.57	.70	.57	70	.79	.79
Permanent	.76†	.94†	.74	.91	.92†	.92
(Instrument: c_1)	(.027)	(.033)	(.032)	(.033)	(.026)	(.029)
Permanent	.68	84	.66	.82	.88	.88
(Instrument: c_{r_1})	(.025)	(031)	(.030)	(.031)	(.032)	(.035)
Permanent	.69	.85	.66	.81	.88	.87
(Instrument: y_1)	(.023)	(.028)	(.028)	(.031)	(.028)	(.03)

Note: See Table 3.

dependence between the *APC* and the permanent income level is therefore built into the model, and an elasticity less than one is the expected result. Thus, a test of the proportionality hypothesis is inappropriate.

Identification of a subsistence level, however, is difficult. The subject has been discussed at length in the Indian literature and the consensus seems to be that an annual income of Rs. 450 per capita, 1970–71 prices (corresponding to Rs. 15-20 per month, 1960–61 prices) adequately describes the subsistence level. If it is assumed that three-year average per capita income y_a adequately reflects permanent status, then $y_a \lesseqgtr$ Rs. 500 should conservatively separate households into subsistence-nonsubsistence categories. If consumption functions are estimated for each group separately, then the prior expectation would be that (a) η' for subsistence households is unity (they are constrained to consume their entire income) and (b) η' for nonsubsistence households is also unity, if the *PIH* is valid.

Table 4 contains the results for the two classifications. For subsistence level households, the unbiased permanent elasticity is very close to one, and never significantly different from one. The most interesting result regarding these households is not that the elasticity is equal to one, but rather that the permanent *marginal* propensity to consume income is not different from one. These results conform very well to any prior expectation about subsistence level behavior. Subsistence households, if properly defined, *should* have a marginal propensity equal to one. Indeed, my results suggest an alternative definition of a subsistence level; in particular, that such a level (or range) is one after which the permanent marginal propensity becomes less than one.

The results for nonsubsistence households represent the strongest possible test of the *PIH*. The a priori dependence between the propensity to consume and permanent income has been "purged" from the data. The results, however, are not favorable to the proportionality assumption. For *all* instruments (arithmetic and logarithmic models), the elasticities are significantly less than one at the 1 percent level of confidence. These elasticities

range from .87 to .92, *log* model. The preferred result $\eta\dagger$ yields η' equal to .92 which is significantly less than unity, but is also considerably larger than the corrected Keynesian elasticity, $\eta^* = .79$.

IV. Conclusions

This paper extends Friedman's permanent income model by explicitly allowing for the distinction between pure measurement errors and transitory terms in the observed variables. Incorporation of this distinction in the theoretical (and empirical) framework is necessary for valid, and direct, tests of the permanent income hypothesis. These tests do not need the measurement error associated with each individual's income per se; rather, only the variance of these errors is necessary.

This extended model not only allows for proper tests of the *PIH* (i.e., those that do not incorporate assumptions additional to the *PIH*) but also makes possible a correct comparison between the traditional and permanent income theory of consumer behavior. Neither hypothesis is supported *in toto.* The contention of the *PIH* that measured consumption elasticities are downwardly biased estimates of the true permanent elasticities is supported by the data—the difference between the two is large and in a direction predicted by the *PIH*, i.e., $\eta' > \eta^*$. However, a major and controversial aspect of the *PIH* is strongly rejected by the data—the elasticity between permanent components is less than unity. This result is subject to all the usual reservations about the applicability of a theory developed in the West for the households of rural India. However, if the permanent income hypothesis is general in its construction (and it is), then these results constitute one of the few, relatively unambiguous, refutations of the proposition that consumption is proportional to permanent income.

Separate consumption functions were estimated for subsistence and nonsubsistence households. The latter set of households were found to have not only a lower consumption elasticity but also a lower permanent marginal propensity to consume. It is an oft-noted result that the *APC* and *MPC* decline with increases in current income. The observed decline of *MPC* out of *permanent* income suggests that income redistribution policies are likely to have, *ceteris paribus*, a negative impact on the supply of household savings, and consequently growth, in the *LDC*s.

The importance of the bias caused by measurement errors is revealed at all stages of the analysis. This is not surprising given the magnitude of the error variances that are observed. These variances are found to be 43 and 26 percent of income variance (arithmetic and multiplicative errors, respectively) for the first year of the survey. This rather large magnitude suggests that researchers should be more cautious than usual with one-shot survey data, especially if these data are collected in rural areas. The error variances are observed to drop radically (to about 8–10 percent) for the second and third years of the survey, thereby suggesting that there is a "learning by doing" aspect to the collection and responses of survey data.

As a by-product, the results of this paper indicate that the length of the consumer's horizon is closer to Friedman's three-year estimate than the short horizons observed by Holbrook. The horizon was found to be between two and three years, but certainly not greater than three.

Appendix A: Data and Definitions

Data. The data are based on a panel survey of 4,118 households in rural India, 1968–69 to 1970–71. The survey, known as the *Additional Rural Income Survey,* was conducted by the National Council for Applied Economic Research (*NCAER*), New Delhi. The survey over-sampled high-income households and gathered data primarily on the pattern of income, consumption and savings of rural households.

For purposes of analysis, only households that were cultivators (self-cultivation on owned or leased land greater than .05 acres) for all three years of the survey were selected. This reduced the sample size from 4,118 to 2,532. Further, households with negative incomes and/or savings (change in net worth) that were estimated to be *greater* than income

for any year of the survey were excluded from analysis. This elimination reduces the sample to 2,453 households.

Definitions:

(a) *Income y*—The income of a household is defined as the total of the earnings of all the members of a household during a reference period. This income can be business income (farm or otherwise), wages, rents (land and house property), interest and dividends on financial investments, and pension and regular contributions.

(b) *Savings s*—The savings of a household is defined as the change in its net worth. This figure is adjusted for capital transfers. In other words, household savings s is defined to be $s = dA - dL - dK$, where dA = gross change in the value of physical and financial assets, dL = net change in liabilities, and dK = net *inflow* of capital transfers. Consumer durables and nonmonetized investment are included in dA. Savings in the form of currency or gold and silver are not included due to lack of reliable data; nor has any adjustment been made for capital gains or losses incurred by the household. Depreciation on assets is also ignored.

(c) *Consumption c*—In addition to the data on income and savings, the *NCAER* survey collected *independent* information on the consumption of households. About twenty-five food items, ten nonfood items (fuel, clothing, medicines, etc.), and "other" items (marriages, funerals, unexpected travel, etc.) comprise the information on consumption expenditures. Addition of these items yields one estimate of consumption c. Subtraction of savings from income yields a residual estimate of c, $c_r = y - s$.

Appendix B: Derivation of Measurement Error Variances

Measurement errors are presumed to exist in all three variables: consumption, savings, and income. These errors are assumed to be related either additively to the true values (arithmetic model) or multiplicatively (logarithmic model). A "method of moments" approach is used to derive the respective error variances; the notation is identical to that used in the text. The reader is referred to the author's working paper for a more detailed derivation.

Arithmetic Model: Errors are assumed to be distributed independently of the true values, of each other, and to have zero mean. Thus,

$$\text{(A1)} \quad z = z^* + z^0 \qquad z = c, s, y$$

$$\text{(A2)} \quad cov(z^*, z^0) = 0$$

$$\text{(A3)} \quad E(z^0) = 0; \; E(z) = E(z^*)$$

As discussed in the text, two kinds of errors are present in consumption; a systematic underestimation by a factor α, and a random error c^0. Consequently (A1) is modified for consumption as

$$\text{(A1}'\text{)} \quad c = \alpha c^* + c^0, \qquad 0 < \alpha < 1$$

The real values are related by the identity

$$\text{(A4)} \quad y^* = c^* + s^*$$

Given assumptions (A1) and (A2), the observed variances and covariances are

$$\text{(A5a)} \quad var\ z = var\ z^* + var\ z^0 \qquad z = y, s$$

$$\text{(A5b)} \quad var\ c = \alpha^2\ var\ c^* + var\ c^0$$

$$\text{(A5c)} \quad cov(s, y) = cov(s^* + s^0, y^* + y^0) = cov(s^*, y^*)$$

$$\text{(A5d)} \quad cov(c, y) = cov(\alpha c^* + c^0, y^* + y^0) = \alpha\ cov(c^*, y^*)$$

(Strictly speaking, probability limits should be used for representing observed variables in (A5). For notational convenience, this usage is suppressed.)

Equations (A1)–(A5) and particularly the use of the identity $y^* = c^* + s^*$ are sufficient to isolate the error variances. In terms of observed values, they are

$$\text{(A6a)} \quad var\ y^0 = var\ y - cov(s, y) - \alpha^{-1}\ cov(c, y)$$

$$\text{(A6b)} \quad var\ s^0 = var\ s + \alpha^{-1}\ cov(c, s) - cov(s, y)$$

$$\text{(A6c)} \quad var\ c^0 = var\ c + \alpha\ cov(c, s) - \alpha\ cov(c, y)$$

Logarithmic Model: If measurement errors are proportional, rather than additive, the equations (A1)–(A3) are replaced by

(A7) $z = z^* z^0 \quad z = y, s$

(A7′) $c = \alpha c^* + c^0$

(A8) $cov(z^*, z^0) = 0 \qquad z = y, s, c$

(A9) $E(z^0) = 1; \; E(z) = E(z^*)$

If incomes are lognormally distributed, then a *log* transformation of the above equations is preferable. (Equation (A12) does not follow from assumption (A9). The problems posed by this nonequivalence are minor and dealt with later.)

(A10) $Z = Z^* + Z^0 \qquad Z = Y, S$

(A10′) $C = A + C^* + C^0 \quad (A = ln\, \alpha)$

(A11) $cov(Z^*, Z^0) = 0$

(A12) $E(Z^0) = 0$

Estimation of *var* Y^0, etc. is desirable to remove the bias from regressions involving the *log* transformations of variables (for example, equation (6)). Though equations (A10)–(A12) are identical to equations (A1)–(A3), a parallel, straightforward solution of *var* Z^0 does not exist. This is due to the fact that the identity $y^* = c^* + s^*$ does not exist in the *log* transformation, i.e., $Y^* \neq C^* + S^*$. An indirect procedure is employed instead. The parameters of the distribution of z^0 are derived first; the *log* transformation of these distributions yield estimates of *var* Z^0.

The multiplicative errors of equation (A7) are assumed to be distributed independently of the true values. According to a theorem by Leo Goodman,

(A13) $$var\, z = [E(z^0)]^2\, var\, z^* + [E(z^*)]^2\, var\, z^0 + var\, z^*\, var\, z^0$$

Equations (A4), (A5c), and (A5d) also hold for the multiplicative error assumption. The system of equations can now be solved for *var* z^0, and results in

(A14) $$var\, y^0 = \frac{\alpha\, var\, y - cov(c, y) - \alpha\, cov(s, y)}{\alpha[E(y)]^2 + cov(c, y) + \alpha\, cov(s, y)}$$

(A15) $$var\, c^0 = \frac{var\, c - \alpha\, cov(c, y) + \alpha\, cov(c, s)}{[E(c)]^2 + \alpha\, cov(c, y) - cov(c, s)}$$

(A16) $$var\, s^0 = \frac{\alpha\, var\, s + cov(c, s) - \alpha\, cov(s, y)}{\alpha[E(s)]^2 - cov(c, s) + \alpha\, cov(s, y)}$$

These variables are derived under the assumption that errors z^0 are distributed with mean one, rather than lognormally distributed with mean zero. The established relationship between normal and lognormal distributions (see J. Aitchison and J.A.C. Brown) allows the derivation of *var* z^0. In particular, if $W = log\, w$ is normally distributed as $N(\mu, \sigma^2)$, then w is lognormally distributed as $N'(\mu', \sigma^{2'})$, where

(A17a) $\mu' = \exp\,(\mu + 0.5\sigma^2)$

(A17b) $\sigma^{2'} = \exp\,(2\mu + \sigma^2)\,(\exp\,(\sigma^2) - 1)$

This relationship, the assumption $E(z^0) = 1$, and some algebra yields expressions for σ^2 or *var* Z^0:

$$\sigma^2 = ln\,(1 + \sigma^{2'})$$

where $\sigma^{2'} = var\, z^0$, and expressions for *var* z^0, in terms of observed parameters, are as given in equations (A14)–(A16).

REFERENCES

J. Aitchison and J. A. C. Brown, *The Lognormal Distribution*, Cambridge 1957.

S. S. Bhalla, "Measurement Errors and the Permanent Income Hypothesis," R-2132-RF, Rand Corp. 1976.

Milton Friedman, *A Theory of the Consumption Function*, Princeton 1957.

———, (1963a) "Note on Nissan Liviatan's Paper," in Carl Christ, ed., *Measurement in Economics*, Stanford 1963.

———, (1963b) "Windfalls, the 'Horizon,' and Related Concepts in the Permanent-Income Hypothesis," in Carl Christ, ed., *Measurement in Economics*, Stanford 1963.

L. Goodman, "On the Exact Variance of Products," *J. Amer. Statist. Assn.*, Dec. 1960, *60*, 708–13.

R. Holbrook, "The Three-Year Horizon: An Analysis of the Evidence," *J. Polit. Econ.*, Oct. 1967, *75*, 750–54.

M. Landsberger, "Consumer's Discount Rate and the Horizon: New Evidence," *J. Polit. Econ.*, Nov./Dec. 1971, *79*, 1346–59.

N. Liviatan, "Tests of the Permanent Income Hypothesis Based on a Reinterview Saving Survey," in Carl Christ, ed., *Measurement in Economics*, Stanford 1963.

Thomas Mayer, *Permanent Income, Wealth and Consumption*, Berkeley 1972.

F. Modigliani and R. Brumberg, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in Kenneth K. Kurihara, ed., *Post Keyensian Economics*, New Brunswick 1954.

K. Mohabatt and E. O. Simos, "Consumer Horizon: Further Evidence," *J. Polit. Econ.*, Aug. 1977, *85*, 851–58.

A. Zellner, "Tests of Some Basic Propositions in the Theory of Consumption," *Amer. Econ. Rev. Proc.*, May 1960, *50*, 565–73.

National Council of Applied Economic Research, *Additional Rural Incomes Survey*, New Delhi 1974.

Approximating Expected Utility by a Function of Mean and Variance

By H. Levy and H. M. Markowitz*

Suppose that an investor seeks to maximize the expected value of some utility function $U(R)$, where R is the rate of return this period on his portfolio. Frequently it is more convenient or economical for such an investor to determine the set of mean-variance efficient portfolios than it is to find the portfolio which maximizes $EU(R)$. The central problem considered here is this: would an astute selection from the E,V efficient set yield a portfolio with almost as great an expected utility as the maximum obtainable EU?

A number of authors have asserted that the right choice of E,V efficient portfolio will give precisely optimum EU if and only if all distributions are normal or U is quadratic.[1] A frequently implied but unstated corollary is that a well-selected point from the E,V efficient set can be trusted to yield almost maximum expected utility if and only if the invector's utility function is approximately quadratic, or if his a priori beliefs are approximately normal. Since statisticians frequently reject the hypothesis that return distributions are normal, and John Pratt and Kenneth Arrow have each shown us absurd implications of a quadratic utility function, some writers have concluded that mean-variance analysis should be rejected as the criterion for portfolio selection, no matter how economical it is as compared to alternate formal methods of analysis.

Consider, on the other hand, the following evidence to the contrary. Suppose that two investors, let us call them Mr. Bernoulli and Mr. Cramer, have the same probability beliefs about portfolio returns in the forthcoming period; while their utility functions are, respectively,

$$U(R) = log(1 + R) \tag{1}$$

$$U(R) = (1 + R)^{1/2} \tag{2}$$

Suppose that Mr. Cramer and Mr. Bernoulli share beliefs about exactly 149 portfolios. In particular suppose that the 149 values of E, V, $Elog(1 + R)$ and $E(1 + R)^{1/2}$ they share happen to be the same as that of the annual returns observed for 149 mutual funds during the period 1958 through 1967, as reported below. (We are not necessarily recommending unadjusted past data as predictors of the future; rather we are using these observations as one example of "real world" moments.)

Now let us suppose that Mr. Bernoulli, having read William Young and Robert Trent, decides that when he knows the E and the V (or the standard deviation σ) of a distribution he may guess its expected utility to him by the formula:

$$Elog(1 + R) \simeq (log(1 + E + \sigma) + log(1 + E - \sigma))/2 \tag{3}$$

He would find that there is a .995 correlation between the pairs (actual mean $\log(1 + R)$, estimated mean $\log(1 + R)$) for the 149 such pairs provided by the 149 historical distributions. Furthermore, the regression relation (over the sample of 149) between the actual mean $log(1 + R)$ and the estimate provided by (3) is

$$\text{actual} = 0.002 + 0.996 \cdot \text{estimated} \tag{4}$$

As it happens, the portfolio which maximized the approximation (3) also maximized the expected value of the true utility (1). If Mr.

*Jerusalem School of Business Administration, Hebrew University, and Thomas J. Watson Research Center, IBM Corporation, respectively We have benefited greatly from technical assistance and the helpful comments of Y. Kroll, and from discussions with Marshall Sarnat

[1]Analyses of relationships between E, V efficiency, on the one hand, and quadratic utility and/or normal distributions, on the other hand, may be found for example in James Tobin (1958, 1963), Markowitz, Martin Feldstein, Giora Hanoch and Levy (1969, 1970), and John Chipman.

Bernoulli selected among the 149 portfolios on the basis of (3) he would, in this instance, do precisely as well as if he had used the true criteria (1). Finally, as will be shown later, (3) always increases with E and decreases with σ, thus is always maximized by an E,V efficient portfolio.

Mr. Cramer, seeing the good fortune of Mr. Bernoulli in finding an approximation to his expected utility based on E and V alone, might try the corresponding approximation to his own utility function, namely:

$$(5)\quad EU \simeq (U(1 + E + \sigma) + U(1 + E - \sigma))/2$$

where U is now given by equation (2). Mr. Cramer would be delighted to find that the correlation between predicted and actual for his utility function is .999; the regression relationship is

$$(6)\quad \text{actual} = -.013 + 1.006 \cdot \text{estimated}$$

The portfolio, among the 149, which maximized the approximation (5) also maximized the true expected utility (2); and (2) is always maximized by a portfolio in the E,V efficient set.

Suppose that a third investor, a Mr. X, does not know his current utility function—has just not taken the time recently to analyze it as prescribed by John von Neumann and Oskar Morgenstern—but does know that equation (5) provides about as good an approximation to his utility function as it does to those of Mr. Bernoulli and Mr. Cramer. He also knows, from certain properties which he is willing to assume concerning his utility function, that equation (5) is maximized by an E,V efficient portfolio. If Mr. X can carefully pick the E,V efficient portfolio which is best for him, then Mr. X, who still does not know his current utility function, has nevertheless selected a portfolio with maximum or almost maximum expected utility.

In this paper we present a class of approximations $f_k(E, V, U(\cdot))$ where $k \geq 0$ is a continuous parameter distinguishing one method of approximation from another. For $k = 1$ we get equation (5); for $k = 0$ we have a method proposed by Markowitz. We shall examine some empirical relationships between EU and $f_k(E,V,U(\cdot))$ for various utility functions, empirical distributions, and values of k. We shall explain these empirical results in terms of a simple analysis of the expected difference between a utility function and an approximating function. We shall also consider certain objections to E, V analysis, due to Karl Borch, Pratt, and Arrow in light of our empirical results, our analysis of expected difference, and a reconsideration of Pratt's analysis of risk aversion for the kinds of quadratic approximations we use.

I. A Class of Approximations

Markowitz used two methods to approximate EU by a function $f(E, V)$ depending on E and V only. The first is based on Taylor-series around zero:

$$(7)\quad U = U(0) + U'(0)R + .5U''(0)R^2 \ldots$$

hence

$$(8)\quad EU \simeq U(0) + U'(0)E + .5U''(0)(E^2 + V)$$

The second approximation is based on a Taylor-series around E:[2]

$$(9)\quad U = U(E) + U'(E)(R - E) + .5U''(E)(R - E)^2 \ldots$$

hence

$$(10)\quad EU \simeq U(E) + .5U''(E)V$$

In tests with empirical distributions and the logarithmic utility function (by Markowitz, and by Young and Trent) the approximation in (10) performed markedly better than the approximation in (8).

Both approximations involve fitting a quadratic to $U(R)$ based on properties of U (i.e., U, U', and U'') at only one value of R ($=0$ or E, respectively). The present authors conjectured that a better approximation perhaps could be found by fitting the quadratic to three judiciously chosen points

[2]A somewhat different use of this Taylor-series to justify mean-variance analysis is presented by S. C. Tsiang. See also Levy regarding the Tsiang analysis.

on $U(R)$. To produce a mean-variance approximation the three points must themselves be functions of at most E, V and the function $U(\ )$. A class of such functions was selected in which the quadratic was fit to the three points

$$(11)\quad (E - k\sigma, U(E - k\sigma)), (E, U(E)), (E + k\sigma, U(E + k\sigma))$$

The quadratic passing through these three points can be written as

$$(12)\quad Q_k(R) = a_k + b_k(R - E) + c_k(R - E)^2$$

To simplify notation we will often write Q, a, and b for Q_k, a_k, and b_k, the subscript k being understood. Equation (12) implies

$$(13)\quad EQ = a + cV$$

Solving

$$(14)\quad U(E - k\sigma) = a + b((E - k\sigma) - E) + c((E - k\sigma) - E)^2 = a - bk\sigma + ck^2\sigma^2$$

$$U(E) = a + b0 + c0^2$$

$$U(E + k\sigma) = a + bk\sigma + ck^2\sigma^2$$

we find that

$$(15)\quad a = U(E)$$

$$b = \frac{U(E + k\sigma) - U(E - k\sigma)}{2k\sigma}$$

$$c = \frac{U(E + k\sigma) + U(E - k\sigma) - 2U(E)}{2k^2\sigma^2}$$

hence

$$(16)\quad f_k(E, V, U(\cdot)) = EQ = U(E) + \frac{U(E + k\sigma) + U(E - k\sigma) - 2U(E)}{2k^2}$$

If we substitute $k = 1$ and simplify we obtain equation (5). If we define the approximation in (10) as f_0, and let $k \to 0$ in (16) we find that[3]

$$(17)\quad f_k \to f_0 \text{ as } k \to 0$$

For given k and U, (16) clearly depends on only E and V. It is not immediately clear that f is always maximized by an E,V efficient portfolio. Each of the utility functions used in our experiments has $U' > 0$, $U'' < 0$, and $U''' \geq 0$ for all rates of return $R > -1.0$. These three properties are sufficient to assure us that f is maximized by an E,V efficient portfolio, provided that $E - k\sigma > -1$ for all portfolios considered.[4]

II. Analysis of Error Functions

For a given k, and for any probability distribution for which the specified moments exist, the difference between EU and $f(E, V, U(\cdot))$ may be written as

$$(18)\quad D_k = EU - EQ_k = Ed_k(R; E, V, U(\cdot))$$

where Q_k is given in (12) and

$$(19)\quad d_k(R) = U(R) - Q_k(R)$$

[3]That $f_k \to f_0$ as $k \to 0$ follows readily if we show that

$$\frac{U(E + k\sigma) + U(E - k\sigma) - 2U(E)}{k^2} \to U''(E)V$$

This may be shown by computing the second derivative of the numerator with respect to k, the second derivative of the denominator, and applying L'Hospital's rule.

[4]Differentiating (16) with respect to σ we find, since $U'' < 0$, that

$$\frac{\partial f_k}{\partial \sigma} = \frac{U'(E + k\sigma) - U'(E - k\sigma)}{2k} < 0$$

Differentiating (16) with respect to E, and substituting $\xi = k\sigma$, we find that

$$\text{(a)}\quad \partial f_k/\partial E = U'(E) + [U'(E + \xi) + U'(E - \xi) - 2U'(E)]V/2\xi^2$$

For positive or negative η Taylor's theorem implies

$$\text{(b)}\quad U'(E + \eta) = U'(E) + U''(E)\eta + .5U'''(E + \theta)\eta^2$$

for some θ between 0 and η. Substituting (b) for $U'(E + \xi)$ and $U'(E - \xi)$ in (a), and using $U' > 0$ and $U''' \geq 0$, we find (for some θ_1 and θ_2 between 0 and ξ) that $\partial f_k/\partial E = U'(E) + [U'''(E + \theta_1) + U'''(E - \theta_2)]V/4 > 0$.

TABLE 1—$d_k(R) = U(R) - Q_k(R)$ FOR $U = log_e(1 + R)$; $E = .1$; $k = 0$ or 1; AND $\sigma = .15$

R	U(R)	$Q_0(R)$	$U - Q_0$	$Q_1(R)$	$U - Q_1$
.70	-1.20397	-.89643	-.30755	-.90348	-.30049
.50	-.69315	-.59891	-.09424	-.60373	-.08942
. .					
.30	-.35667	-.33444	-.02223	-.33735	-.01933
.25	-.28768	-.27349	-.01419	- 27596	-.01172
.20	-.22314	-.21461	-.00854	- 21667	-.00648
.15	- 16252	-.15779	-.00473	-.15946	-.00306
.10	-.10536	-.10304	-.00232	-.10433	-.00103
.05	-.05129	-.05035	-.00094	-.05129	.00000
00	.00000	.00027	-.00027	-.00034	00034
.05	.04879	.04882	-.00003	.04853	.00026
.10	.09531	.09531	-.00000	.09531	-.00000
.15	.13976	.13973	.00003	.14001	- 00024
.20	.18232	.18209	.00023	.18262	-.00030
.25	.22314	.22238	.00077	22314	.00000
.30	.26236	.26060	.00176	26158	.00078
.35	.30010	.29676	.00335	29794	.00217
.40	.33647	.33085	.00563	33221	00427
.45	.37156	.36287	.00869	.36439	.00718
.50	.40546	.39283	01263	.39449	.01098
55	.43825	.42072	.01753	.42250	.01576
60	.47000	.44655	.02345	44842	.02158
. .					
00	.69315	.57878	.11437	.58075	11240
.50	.91629	.55812	.35817	.55846	.35783
.00	1.09861	.33086	.76775	.32761	.77100
50	1.25276	-.10301	1.35577	-.11179	1.36455
00	1.38629	-.74349	2.12978	-.75975	2.14604

'or example, Table 1 presents $d_k(R)$ for $J = log(1 + R)$, for $k = 0$ and $k = 1$, for $\bar{z} = .1$, for $\sigma = .15$, and for various values of ?. Much about the joint distribution of EU nd $f_k(E, V, U(\cdot))$ is explained by such ables plus general properties of the distribu-ions involved. Consider the fourth column of 'able 1 showing $d_k(R)$ for $k = 0$. Among the 49 mutual fund distributions mentioned :arlier, those with E in the neighborhood of 10 all have every year's return between a 30 ercent loss and a 60 percent gain for the 'ear. (For example, 64 distributions had $.08 \leq E \leq .12$; all were within the range indi-:ated.) $d_0(R)$ goes from $-.022233$ at $R = -.3$, to $+.023454$ at $R = .6$, with substan-ially smaller values of $|d_0|$ in between. If we magine spreading a probability distribution hroughout the interval $-.3$ to $+.6$, keeping $\bar{z} = .1$, there is a limit to how large we can nake $|Ed_0|$. In fact, if we assume that the listribution can take on only the values listed n the table ($-.3$ to $+.6$ by steps of .05), then ι little linear programming will show us that the worst distribution, the one with the greatest $|E(d)|$, is one with a probability of 3/8 of $R = -.3$, a probability of 5/8 of $R = +.35$, and with $E(d) = -.00649$. None of the 149 historical distributions had this worst possible distribution. Insofar as they were less skewed, positive errors above E tended to cancel negative errors below E. Insofar as they clustered closer to the mean than in our worst case, the absolute value of the deviations were smaller.

If we recomputed Table 1 for some other E, say $E = .15$, the new table would appear very much like the old. The principal difference in column 4 would be that the smallest values of $|d_0(R)|$ would be centered around the new mean, $R = .15$. An analysis similar to that above would again explain why $|Ed|$ was small for those historical distributions which had E in this new neighborhood.

We should examine the concept of $|Ed|$ being small. A statement to the effect that the difference between EU and EQ is less than some number (like .0065) is, in itself, of absolutely no value either in explaining the correlation between EU and EQ, or in judging whether Q is a good approximation to U in practice. For a utility function is only defined up to an arbitrary choice of unit and scale. In particular, if we multiply U by an arbitrary positive constant, obtaining a precisely equivalent utility function, we also multiply by the same constant the approximations (8), (10), and (16), and therefore multiply by the same constant the difference between EU and each of these. Thus we can make $|Ed|$ arbitrarily close to zero by using the utility function $V = bU(R)$ for sufficiently small b.

The arbitrary choice of unit and scale, however, does not change certain measurements: the correlation and regression coefficients are unaffected; and the appearance of a comparison between the plot of U against R vs. a plot of Q against R is also unaffected in the following sense. Suppose that we plot U in Table 1 against R for $R = -.3$ to $+.6$. Since U rises from about $-.36$ to about $+.47$ we might allow .1 utility units per inch of vertical scale. If we also plot on the same graph $Q_0(R)$ from the third column of the table, Q would be two-tenths of an inch above U at

$R = -.3$, about two-tenths of an inch below U at $R = .6$; but for much of their lengths the two curves would be virtually indistinguishable as they rose from the lower left to the upper right-hand corner of the page. Suppose now that we change the origin and scale of the utility measure. If we still want the two curves to fill the page we simply relabel the vertical axis leaving the two curves unchanged.

If we define σ_{Ed} to be the standard deviation of Ed over a set of distributions, where Ed is the mean value of d for a given distribution, and similarly define $\sigma_f = \sigma_{EQ}$ as the standard deviation of f over the set of distributions, then it can be shown[5] that the correlation $\rho_{EU,f}$ between EU and f over the set of distributions is at least

$$\rho_{EU,f} \geq (1 - \gamma^2)^{1/2} \tag{20}$$

where

$$\gamma = \sigma_{Ed}/\sigma_f \tag{21}$$

Thus if f is ten times as variable as Ed then $\rho_{EU,f}$ is at least $(.99)^{1/2} = .995$.

Column 6 of Table 1 presents $d_1(R)$. This equals zero at $R = E - \sigma$, $R = E$, and $R = E + \sigma$ as planned. Just considering $-.3 \leq R \leq .6$, the f_1 approximation is definitely superior to f_0 in the range from $R = -.05$ through $-.3$, and from $+.25$ to $.6$. On the other hand, f_0 fits better near the mean, i.e., from .00 through .20 among the values listed. The empirical results presented in the following sections indicate, for the distributions and utility functions explored, whether it was more beneficial to approximate a bit better near the mean, as does f_0, or to hold up well over wider range as does f_1.

Table 1 also shows values of $d_0 = U - Q_0$ and $d_1 = U - Q_1$ for more extreme values of $|R - E|$ than discussed thus far. We see for example that, for an investor with a logarithmic utility function, f_0 is likely to be a poor approximation to EU for a distribution with $E = .1$ and with nontrivial probabilities of, say, $R = -.7$ and $R = 1.5$. More generally, the empirical results reported below would be less favorable to mean-variance approximations if we were dealing with much more speculative distributions. The mean-variance analysis is thus more suitable for investor's and opportunity sets in which such extremes have very low probabilities in the portfolios which maximize both EU and f_k.

III. Empirical Results

For annual returns of 149 investment companies, 1958–67,[6] Table 2 shows the correlation between $EU(R)$ and f_k $(E, V, U(\cdot))$ for $k = .01, .1, .6, 1.0$, and 2.0 and for various utility functions. (We also computed correlation coefficients for a few other values of the a and b coefficients in the utility functions, with results which one might expect by interpolating or extrapolating the results reported in Table 2. For example, the exponential utility with $b = 20$ was even more of an exception to the general rule than is the case with $b = 10$ reported here.) A row of Table 2 presents correlation ρ as a function of k, for some given utility function. In every case reported in Table 2, with the exception of the exponential utility function with $b = 10$, ρ is a nonincreasing function of k; hence $\rho_{01} \geq \rho_k$ for all k. Note ρ_k, as a function of k, is frequently quite flat between $k = .01$ and 1.0, but drops faster from $k = 1$ to $k = 2$. We did not calculate the ρ_{00} correlations but, from continuity considerations, we assume

[5]From (19) and $f = EQ$ it follows that

$$\rho_{EU,f} = cov\,(f, f + Ed)/(\sigma_f \sigma_{f+Ed}) \tag{a}$$

A short calculation shows that $cov\,(f, f + Ed) = var\,(f) + \rho_{f,Ed}\sigma_f\sigma_{Ed} = var\,(f)(1 + \gamma\rho_{f,Ed})$, and that $\sigma_{f+Ed} = \sigma_f(1 + \gamma^2 + 2\gamma\rho_{f,Ed})^{1/2}$ Substituting these two formulas into (a) we get

$$\rho_{EU,f} = (1 + \gamma\rho_{f,Ed})/(1 + \gamma^2 + 2\gamma\rho_{f,Ed})^{1/2} \tag{b}$$

For $\gamma < 1$ this is a continuous, positive function of $\rho_{f,Ed}$ in the range $-1 \leq \rho_{f,Ed} \leq +1$. If we differentiate (b) with respect to $\rho_{f,Ed}$, set the resulting expression equal to zero, assume $0 < \gamma < 1$, and solve for $\rho_{f,Ed}$, we find that $\partial\rho_{EU,f}/\partial\rho_{f,Ed} = 0$, only at $\rho_{f,Ed} = -\gamma$. Substituting this into (b) we find that at this stationary point $\rho_{EU,f} = (1 - \gamma^2)^{1/2}$. We may confirm that is a minimum rather than a maximum or inflection point by noting, from (b), that $\rho_{EU,f} = +1 > (1 - \gamma^2)^{1/2}$ for $\rho_{f,Ed} = \pm 1$.

[6]The annual rate of return of the 149 mutual funds are taken from the various annual issues of A. Wiesenberger and Company. All mutual funds whose rates of return are reported in Wiesenberger for the whole period 1958–67 are included in the analysis.

TABLE 2—CORRELATION BETWEEN $EU(R)$ AND $f_k(E, V, U(\cdot))$ FOR ANNUAL RETURNS OF 149 MUTUAL FUNDS, 1958–67

Utility Function		k = 0.01	0.1	0.6	1.0	2.0
Log(1 + R)		0.997	0.997	0.997	0.995	0.983
$(1+R)^a$	a=0.1	0.998	0.998	0.997	0.997	0.988
	a=0.3	0.999	0.999	0.999	0.998	0.995
	a=0.5	0.999	0.999	0.999	0.999	0.998
	a=0.7	0.999	0.999	0.999	0.999	0.999
	a=0.9	0.999	0.999	0.999	0.999	0.999
$-e^{b(1+R)}$	b=0.1	0.999	0.999	0.999	0.999	0.999
	b=0.5	0.999	0.999	0.999	0.999	0.999
	b=1.0	0.997	0.997	0.997	0.996	0.995
	b=3.0	0.949	0.949	0.941	0.924	0.817
	b=5.0	0.855	0.855	0.852	0.837	0.738
	b=10.	0.447	0.449	0.503	0.522	0.458

that they are close to those found for $k = .01$. For most cases considered $\rho_{01} > .99$.

The correlations for the exponential with $b = 10$ are much lower than those of the other utility functions reported in Table 2. In our 1977 paper we analyze this utility function at a greater length than space permits here, and arrive at two conclusions. The first conclusion is that an investor who had $-e^{-10(1+R)}$ as his utility function would have some very strange preferences among probabilities of return. Reasonably enough, he would not insist on certainty of return. For example, he would prefer (a) a 50–50 chance of a 5 percent gain vs. a 25 percent gain rather than have (b) a 10 percent gain with certainty. On the other hand there is no R which would induce the investor to take (a) a 50-50 chance of zero return (no gain, no loss) vs. a gain of R rather than have (b) a 10 percent return with certainty. Thus a 50-50 chance of breaking even vs. a 100 percent, or 300 percent, or even a 1000 percent return, would be considered less desirable than a 10 percent return with certainty. We believe that few if any investors have preferences anything like these. A second conclusion, more important than the first as far as the present discussion is concerned, is that even if some unusual investor did have the utility function in question, if he looked at his $d_k(R)$ in advance he would be warned of the probable inapplicability of mean-variance analysis. The corresponding version of Table 1 (scaled to have about the same σ_f for the two approximations, as (20) suggests) shows d_k to generally be more than an order of magnitude greater for the exponential with $b = 10$ than for the logarithmic utility function.

TABLE 3—CORRELATION BETWEEN $EU(R)$ AND $f_{01}(E, V, U(\cdot))$ FOR 3 HISTORICAL DISTRIBUTIONS

Utility Function		Annual returns on 97 stocks[8]	Monthly returns on 97 stocks[8]	Random portfolios of 5 or 6 stocks[9]
Log(1 + R)		0.880	0.995	0.998
$(1+R)^a$	a=0.1	0.895	0.996	0.998
	a=0.3	0.932	0.998	0.999
	a=0.5	0.968	0.999	0.999
	a=0.7	0.991	0.999	0.999
	a=0.9	0.999	0.999	0.999
$-e^{b(1+R)}$	b=0.1	0.999	0.999	0.999
	b=0.5	0.961	0.999	0.999
	b=1.0	0.850	0.997	0.998
	b=3.0	0.850	0.976	0.958
	b=5.0	0.863	0.961	0.919
	b=10	0.659	0.899	0.768

Table 3 shows the correlation between EU and f_{01} for three more sets of historical distributions. While ρ_k was computed for the same values of k reported in Table 2, we confine our attention to $k = .01$ since, almost without exception, ρ_k was a nonincreasing function of k. The first column of data in Table 3 shows ρ_{01} for annual returns on 97 *U.S.* common stocks during the years 1948–68.[7] It is understood, of course, that mean-variance analysis is to be applied to the portfolio as a whole rather than individual investments taken one at a time. Annual returns on individual stocks were used in this example, nevertheless, as an example of historic distributions with greater variability than that found in the portfolios reported in Table 2. As expected, the correlations are clearly poorer for the individual stocks than they are for the mutual fund portfolios. For $U = log(1 + R)$, for example, the correlation is .880 for the annual returns on stocks as

[7]This data base of 97 *U.S.* stocks, available at Hebrew University, had previously been obtained as follows: a sample of 100 stocks was randomly drawn from the CRSP (Center for Research in Security Prices, University of Chicago) tape, subject to the constraint that all had reported rates of return for the whole period 1948–68. Some mechanical problems reduced the usable sample size from 100 to 97. The inclusion only of stocks which had reported rates of return during the whole period may have introduced selection bias into the sample. It might prove worthwhile to experiment with alternate methods of handling the appearance and disappearance of stocks.

compared to .997 for the annual returns on the mutual funds.

Since monthly returns tend to be less variable than annual returns we would expect the correlations to be higher for the former than the latter. The ρ_{01} for monthly returns on the same 97 stocks are shown in the second column of data in Table 3. For the logarithmic utility function, for example, the correlation is .995 for the monthly returns on individual stocks as compared to .880 for annual returns on the stocks and .997 for annual returns on the mutual funds. On the whole, the ρ_{01} for the monthly returns on individual stocks are comparable to the annual returns on the mutual funds.

Annual returns on individual stocks (i.e., on completely undiversified portfolios) have perceptibly smaller ρ_{01} than do the annual returns on the well diversified portfolios of mutual funds. The third column of data in Table 3 presents ρ_{01} for "slightly diversified" portfolios consisting of a few stocks. Specifically, it shows the correlations between EU and f_{01} on the annual returns for nineteen portfolios of 5 or 6 stocks each randomly drawn (without replacement) from the 97 U.S. stocks.[8] We see that for the logarithmic utility function $\rho_{01} = .998$ for the random portfolios of 5 or 6, up from .880 for individual stocks. The ρ_{01} for the annual returns on the portfolios of 5 or 6 were generally comparable to those for the annual returns on the mutual funds. These results were perhaps the most surprising of the entire analysis. They indicate that, as far as the applicability of mean-variance analysis is concerned, at least for joint distributions like the historical returns on stocks for the period analyzed, a little diversification goes a long way.

In addition to the correlation coefficient, we examine in our 1977 paper other measures of the ability of f_k to serve as a surrogate for EU, and conclude that f_{01} does as well in these comparisons as it does in terms of correlations. For example, we computed the frequency with which any other available portfolio was better than the portfolio which maximized f_k. We found, in particular, that in every case with $\rho > .9$ in Tables 2 or 3 the portfolio with maximum f_{01} was also the portfolio (among the 149 or 97 or 19 considered) with the greatest EU. We cannot say with any precision how high a correlation between f_k and EU is high enough. Be that as it may, for many of the utility functions and distributions considered (chosen in advance as representative of utility functions frequently postulated, or distributions clearly "real world") f_{01} was an almost faultless surrogate for EU. Where f_{01} performed poorly, the user would have been warned in advance by an analysis of expected error.

IV. Some Objections Reconsidered

Since $\rho_{EU,f} < 1$ it can happen that portfolio A has a higher f_k, than portfolio B, while portfolio B has a higher EU. In fact, given any function f of E and V, Borch presents a method for finding distributions A and B such that $f(E_A, \sigma_A) = f(E_B, \sigma_B)$, yet clearly $EU_A > EU_B$ because distribution A stochastically dominates distribution B. (In terms of equation (20), this example has $\gamma = \infty$.)

Borch's argument shows that it is hopeless to seek an f that will be perfectly correlated with EU for all collections of probability distributions. The evidence on the preceding pages nevertheless supports the notion that the imperfect approximations $f_k(E, V)$ are frequently good enough in practice to choose almost optimally among realistic distributions of returns; and the d_k function can be used in advance to judge the suitability of f_k.

The approximation f_k was obtained by fitting the quadratic (12) to three points. We have seen that for certain utility functions and historical distributions of returns, f_k is highly correlated with EU. Pratt and Arrow, on the other hand, have shown that any quadratic utility function had highly undesirable theoretical characteristics. How do we reconcile these two apparently contradictory observations?

[8]We randomly drew 5 stocks to constitute the first portfolio; 5 different stocks to constitute the second portfolio, etc. Since we have 97 stocks in our sample, the eighteenth and nineteenth portfolios include 6 stocks each. Repetition of this experiment with new random variables produced negligible variations in the numbers reported, except for the case of $U = e^{-10(1+R)}$. A median figure is reported in the table for this case.

It is essential here to distinguish between three types of quadratic approximations:

1) Assuming that utility as a function of wealth $V(W)$ remains constant from period to period, a quadratic $q(W)$ is fit to $V(W)$ once and for all. As W changes from time to time, the same $q(W)$ function is used to select portfolios. (Note that V expresses utility as a function of wealth W, in contrast to the previously defined U function which expressed utility as a function of rate of return R. Note also that an unchanging $V(W)$ implies the existence of some $U(R)$ at each point in time, though not necessarily the same $U(R)$ each time. The converse is not true: the existence of a $U(R)$ at each point in time does not necessarily imply an unchanging $V(W)$.)

2) At the beginning of each time period a $q(W)$ function is fit to $V(W)$ at the point W = current wealth; or equivalently $Q(R)$ is fit to $U(R) = V((1 + R)W_{t-1})$ at $R = 0$. Even if $V(W)$ remains constant through time, the quadratic fit $Q(R)$ is changed as wealth changes.

3) The quadratic fit (of $Q(R)$ to $U(R)$) depends on at least E and perhaps σ as well. In this case the fit varies between one distribution and another being evaluated at the same time.

The Pratt and Arrow objections apply to quadratic approximations of type 1.[9] The approximation in (8) is of type 2. Approximations (10) and (16) are of type 3. Each of these three classes of approximations have different risk-aversion properties. We shall confine our attention here to a comparison between the first and third of these. To illustrate the difference between the first and third, we shall compare their "risk premiums" for small risks as defined by Pratt.

Pratt's results may be expressed as follows. Let $Pr_i(R)$ $i = 1, 2, 3, \ldots$ be a sequence of probability distributions such that each has the same mean:

$$\text{(22a)} \quad \int R\,dPr_i(R) = E_0 \quad i = 1, 2, 3, \ldots$$

and such that their standard deviations approach zero:

$$\text{(22b)} \quad \lim \sigma_i = 0$$

where the limit, here and elsewhere in this section unless otherwise specified, is taken as $i \to \infty$. The risk premium for the ith distribution is defined implicitly by

$$\text{(22c)} \quad U(E_0 - \pi_i) = \int U(R)\,dPr_i(R) \quad i = 1, 2, 3, \ldots$$

where the right-hand side equals the expected utility of the ith distribution. Pratt shows that, under certain general conditions,

$$\text{(23)} \quad \lim \pi_i/\sigma_i^2 = -1/2V''((1 + E_0)\,W_{t-1}) / V'((1 + E_0)W_{t-1}) = -1/2U''(E_0)/U'(E_0) = 1/2r((1 + E_0)W_{t-1})$$

where r is defined to be the investor's risk aversion at the wealth associated with return $R = E_0$. Pratt reasonably asserts that π/σ^2, and hence $r((1 + E_0)W_{t-1})$ should be a decreasing function of its one argument, $W = (1 + E_0)W_{t-1}$. He notes that for a quadratic (of type 1) $r(W)$ is an increasing function of W and concludes:

> Therefore a quadratic utility cannot be decreasingly risk-averse on any interval whatever. This severely limits the usefulness of quadratic utility, however nice it would be to have expected utility depend only on the mean and variance of the probability distribution. Arguing "in the small" is no help: decreasing risk aversion is a local property as well as a global one. [p. 132]

That Pratt's conclusion is not correct for an approximation of type 3 will be shown in the following way. If an investor maximized some mean-variance approximation $f(E,\sigma)$, such as f_k for fixed nonnegative k, then his risk premium would be given implicitly by

$$\text{(24)} \quad f(E - \pi, 0) = f(E, \sigma)$$

[9] Pratt correctly asserts that his analysis does not require constant $V(W)$ over time. We must distinguish, however, between a V function varying with time (but not depending on W itself) vs. approximations of type 2 or 3 in which the choice of the quadratic function depends on W or even E and σ. For simplicity, we describe approximations of type 1 in terms of a fixed $V(W)$, rather than try for greater generality here.

Below we define a class of type 3 approximations including, in particular, f_k for all $k \geq 0$. We consider a sequence of probability distributions $Pr_i(R)$ $i = 1, 2, 3, \ldots$ satisfying (22a) and (22b), and find that

$$\text{(25)} \quad \lim \pi_i/\sigma_i^2 = -1/2U''(E_0)/U'(E_0)$$

for all approximations in the class. Thus the risk aversion "in the small" of f_k is precisely the same as that of U itself. We have seen that not all f_k are equally good "in the large." But in the small, they are asymptotically the same as the utility function which they approximate.

We assume that the investor maximizes an approximation $f(E,\sigma)$ which is the expected value of a quadratic of type 3 such that $f = EQ$ in (13) satisfies

$$\text{(26a)} \qquad a = U(E)$$

$$\text{(26b)} \qquad c \to .5U''(E) \text{ as } \sigma \to 0$$

It follows immediately from (10) that f_0 satisfies (26a) and (26b), and from (15) that f_k $k > 0$ satisfies (26a). That f_k also satisfies (26b) follows from L'Hopital's rule applied to the expression for c in (15).

From (13) and (26a), (24), (13) again, and (26b) we infer that

$$\begin{aligned} U(E_0 - \pi) &= f(E_0 - \pi, 0) \\ &= f(E_0, \sigma) \\ &= U(E_0) + c(E_0, \sigma)\sigma^2 \\ &\quad \to U(E_0) \text{ as } \sigma \to 0 \end{aligned}$$

But $U(E_0 - \pi) \to U(E_0)$, and $U' > 0$ throughout, implies

$$\text{(27)} \qquad \lim \pi_i = 0$$

Using Taylor's theorem we may write

$$\begin{aligned} \text{(28)} \quad f(E_0 - \pi, 0) &= U(E_0 - \pi) \\ &= U(E_0) - U'(E_0)\pi \\ &\quad + .5U''(\xi)\pi^2 \end{aligned}$$

where ξ is between $E_0 - \pi$ and E_0. Hence, from (27), $\xi \to E_0$ as $i \to \infty$. Using (28) on the left of (24), (13) on the right of (24), and rearranging terms we get

$$\text{(29)} \quad \pi/\sigma^2 = -c(E_0,\sigma)/(U'(E_0) - .5U''(\xi)\pi)$$

for each distribution in the sequence. This together with (26b) and (27) implies (25).

V. The E,V Investor

Let us return to Mr. X who has not analyzed his utility function recently. Suppose that, when presented with probability distributions of E,V efficient portfolios, he can pick that portfolio which has greater EU than any other E,V efficient portfolio. By definition, his choice will be at least as good as the portfolio which maximizes f_{01}. In addition to the functions of E and V discussed above there may very well be others—not yet explored, or perhaps not yet conjectured—which perform better than does any f_k as a surrogate for EU. Mr. X's choice of portfolio will also be at least as good as the best of all of these in each particular situation.

REFERENCES

Kenneth Arrow, *Aspects of the Theory of Risk Bearing*, Helsinki 1965.

K. Borch, "A Note on Uncertainty and Indifference Curves," *Rev. Econ. Stud.*, Jan. 1969, *36*, 1–4.

J. S. Chipman, "The Ordering of Portfolios in Terms of Mean and Variance," *Rev. Econ. Stud.*, Apr. 1973, *40*, 167–90.

M. S. Feldstein, "Mean Variance Analysis in the Theory of Liquidity Perference and Portfolio Selection," *Rev Econ. Stud.*, Jan. 1969, *36*, 5–12.

G. Hanoch and H. Levy, "The Efficiency Analysis of Choices Involving Risk," *Rev. Econ. Stud.*, July 1969, *36*, 335–46.

——— **and** ———, "Efficient Portfolio Selection with Quadratic and Cubic Utility," *J. Bus., Univ. Chicago*, Apr. 1970, *43*, 181–89.

H. Levy, "The Rationale of the Mean-Standard Deviation Analysis: Comment," *Amer. Econ. Rev.*, June 1974, *64*, 434–41.

——— **and H. Markowitz,** "Mean-Variance Approximations to Expected Utility," T. J. Watson Res. Ctr., IBM, Aug. 1977.

Harry Markowitz, *Portfolio Selection: Efficient Diversification of Investments*, New York 1959; New Haven 1970.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan. 1964, *32*, 122–36.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–87.

———, "The Theory of Portfolio Selection" in Frank H. Hahn and Frank P. R. Brechling, eds., *Theory of Interest Rates*, New York 1963.

S. C. Tsiang, "The Rationale of the Mean Standard Deviation Analysis, Skewness Preference, and the Demand for Money," *Amer. Econ. Rev.*, June 1972, *62*, 354–71.

John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, 3d ed., Princeton 1953.

W. E. Young and R. H. Trent, "Geometric Mean Approximation of Individual Security and Portfolio Performance," *J. Finan. Quant. Anal.*, June 1969, *4*, 179–99.

A. Wiesenberger and Company, *Investment Companies*, New York, annual editions.

The Emergence of Countercyclical U.S. Fertility

By WILLIAM P. BUTZ AND MICHAEL P. WARD*

In recent years the simple positive association between fertility changes and aggregate economic activity in developed countries has been breaking down. Although generalizations often fall victim to events, procyclical fertility had become such a regularity that one analyst called the pattern "one of the most firmly based empirical findings in any of the social sciences."[1] Indeed, before the *U.S.* "baby bust" of the 1960's, many social scientists were predicting continued fertility increases during future economic expansions. In the face of the rapid fertility decline that occurred, however, analysts have appealed to a range of social, institutional, and technological forces that may have swamped the expected rise in fertility: increased contraceptive supply and effectiveness; increases in couples' expected standard of living; women's attitudinal changes toward work and childbearing; government affirmative action programs; and increasing "social awareness" concerning world population growth. Accordingly, both professional and nonprofessional opinion as to the cause of the fertility decline has drifted steadily away from the simple mechanism tying economic well-being to desired fertility.

We suggest an alternative mechanism through which aggregate economic activity affects fertility rates. Our approach is in the spirit of previous work analyzing the microeconomic determinants of fertility.[2] We emphasize the distinction between male and female earnings in affecting fertility and the distinction between families with employed wives and those without. We show that postwar movements in fertility rates—a rapid increase through the 1950's and a sharp decline thereafter—can be reconciled within this framework. Our empirical results indicate that the "baby boom" of the 1950's can be explained as a response to rising male income, whereas the baby bust of the 1960's was due primarily to increases in female wages and income. Our results also indicate that future fertility movements can be expected to move countercyclically.[3]

I. An Empirical Model

Cross-sectional studies of fertility behavior indicate that although male income may be positively related to fertility, increases in female wages have a strong negative effect on birth probabilities.[4] The theoretical underpinnings of these studies derive from a static model of household behavior in which the family is viewed as maximizing utility defined over market goods, leisure, and child services (see fn. 2). Child services are "produced" in the household using as inputs the wife's time and market goods. An increase in the husband's market wage raises family income and, if husband's time is not an important input into the "production" of child services, leads to a higher demand for children. An increase in the wage of an employed woman also adds to family income, but it simultaneously increases the price of children since

*The Rand Corporation and University of California-Los Angeles, respectively. This paper was written under a grant from the National Institute of Child Health and Human Development to the Rand Corporation. We have benefited from useful discussions with Yoram Ben-Porath, John Cogan, Jim Smith, Craig Foch, Marc Nerlove, and Dennis DeTray, and from the comments of the managing editor and an anonymous referee. We are also indebted to Jan Meshkoff, Robert Newman, and Donald Negri for expert research assistance.

[1]See Dorothy Thomas. For underlying research, see Thomas and W. F. Ogburn; Thomas, Virginia Galbraith and Dorothy Thomas; Dudley Kirk (1942, 1960); Hannes Hyrenius; Morris Silver (1965).

[2]The basic formulation is due to Gary Becker; Jacob Mincer; Robert Willis (1971).

[3]These results concern period fertility rates, the number of births in a year per 1,000 women of a specified age. Annual changes in these rates may reflect considerations of completed family size as well as the desired pattern of childbearing over time to reach the target. As with most other studies of time-series fertility, we do not attempt here to estimate the separate contributions of these two factors to observed changes in period fertility rates.

[4]See the various studies in T. Paul Schultz.

the opportunity cost of childbearing and rearing rises at the same time. This lack of symmetry in the treatment of sources of family income stems from the supposition that children are intensive users of the wife's time, in the production function sense.

A further distinction can be drawn between families with and without employed wives.[5] If the wife is employed (and continues to be), then increases in the husband's wage lead to a reduction in her hours of labor market work, but do not alter the price of her time. If the wife is not employed, increases in the husband's wage should increase fertility but by a smaller magnitude. This follows if, as is normally assumed, husbands' and wives' household time inputs are gross substitutes; then increases in the wage of one induce the other to substitute away from market work. Since a nonemployed wife cannot withdraw further from the market, the shadow price of her time and hence the shadow price of children rise when her husband's wage increases. This tends to mitigate the positive effect of an increase in the husband's wage on fertility. Finally, an increase in the wife's market wage that is insufficient to cause her to enter employment will have no effect on the demand for children.

What then are the implications of the static model for period fertility rates? Data on these rates reflect current family decisions about both desired completed family size and the distribution or timing of births over the family's life cycle.[6] If short-run fluctuations in income could be perfectly forecast and if perfect capital markets were available to buffer these fluctuations, then fertility rates would not necessarily move procyclically but only in response to changes in wealth or permanent income. However, to the extent that capital markets are imperfect and current fluctuations are imperfectly forecast, there is an incentive to time consumption so as to more closely match income variations. That is, fertility rates would then tend to move procyclically.

Even in a world of complete certainty and perfect capital markets, timing *is* affected by variations in the price of commodities. If women's wages rise in business upswings, the opportunity cost of her time rises. Couples then have an incentive to postpone the consumption of activities intensive in the use of her time.

We consider the probability B that a couple will have a child in a given year to be dependent on Y_m, the husband's income, the opportunity cost of the wife's time, and a vector of other factors represented by the variable X.[7] For families in which the wife is employed, the opportunity cost of the wife's time is her wage W_f. For nonemployed wives the opportunity cost or shadow price of her time is W_f^*. We take the latter to be a function of husband's income. If B_1 and B_2 are the annual probabilities of a birth for a nonemployed and employed woman, we can write

$$(1)\quad B = \begin{cases} B_1 = B_1^0(Y_m, W_f^*(Y_m), X) \\ \quad = B_1(Y_m, X), \\ \qquad \text{for nonemployed wives} \\ B_2 = B_2(Y_m, W_f, X), \\ \qquad \text{for employed wives} \end{cases}$$

Figure 1(a) describes such a function. Wage movements below the reservation wage W_r have no effect on the probability of a birth, since by definition, the wife would not participate in the labor force at a wage less than W_r.

[5]Willis (1971, 1973) first emphasized this distinction.

[6]Families may be thought to choose target levels of completed fertility based on longer-run considerations (permanent income or wealth, rearing costs, etc.) while timing decisions are affected more by current variables. A complete specification of such a model would also account for the rate of adjustment toward desired numbers of children, and the role of expectations in influencing both current and target fertility. We have ignored these dynamic considerations and instead focused exclusively on the effects of current economic variables on current fertility measures. We are thus unable to identify the separate channels of influence of, for example, current income on permanent income and hence on desired completed fertility, and finally on current fertility rates. We have investigated these dynamic considerations in further work. See the authors.

[7]We ignore nonwage income and the influence of husband's wage on the cost of a child. Implicitly, then, children are costly only in their use of women's time. In using male income rather than wage rates we assume that hours worked by men are exogenous with respect to current fertility rates.

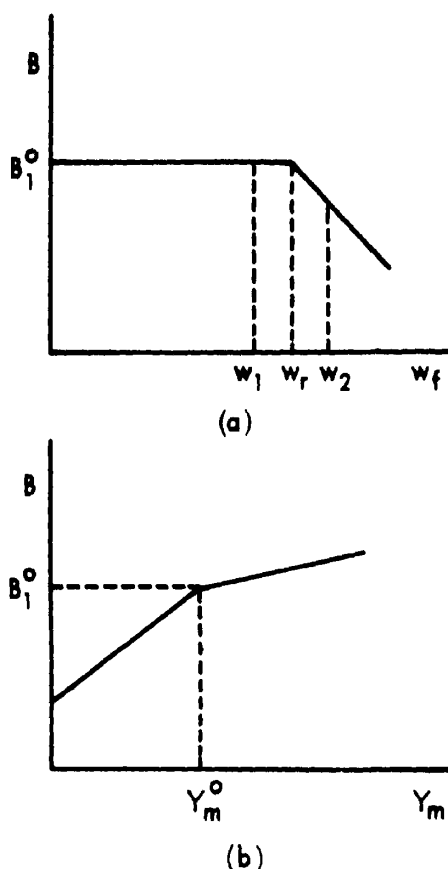



FIGURE 1. BIRTH PROBABILITY AS A FUNCTION OF FEMALE WAGE AND MALE INCOME

For wives with wages above the reservation wage (i.e., those who are employed), increases in the market wage represent increases in the time cost of children, leading to a reduction in the annual birth probability given by the sloped segment in Figure 1(a). Finally, for wives who change their employment status in response to a wage change (say from W_1 to W_2), the change in birth probability is a fraction of that exhibited by women who remained employed.

The change in the average birth probability in response to a wage increase is an average of the responses in these three classes, nonemployed, employed, and transitional, with weights proportional to the fraction of families in these groups. If we let K denote the fraction of families with employed wives, and ΔK the change in that fraction induced by a wage change, then the aggregate response to a change in the wage can be written as

$$(2) \quad dB = K\frac{\partial B_2}{\partial W_f}dW_f + \alpha\Delta K\frac{\partial B_2}{\partial W_f}dW_f$$

The first term is the response within the group of women who remain employed. It represents movement along the sloped segment given in Figure 1(a). The last term is the response of women who change their employment status. It is some fraction α of the response of women who remain employed; the fraction being a function of how close the initial market wage is to the reservation wage for a typical woman in this group. Finally, women who do not enter the work force exhibit no change in their birth probabilities.

The expression ΔK represents the movement into or out of the labor force induced by the change in wage rates. If we denote this by $(\partial K/\partial W_f)dW_f$ then equation (2) can be written as

$$(3) \quad dB = K\frac{\partial B_2}{\partial W_f}dW_f + \alpha\frac{\partial K}{\partial W_f}\frac{\partial B_2}{\partial W_f}(dW_f)^2$$

The same argument can be repeated for changes in male income. In Figure 1(b), we describe the hypothesized response of birth probabilities to changes in male income. Below some level of male income (say Y_m^0), the wife is in the work force. As male income rises within this range, the birth probability rises. Above Y_m^0, the wife is out of the work force and, as argued above, the response of birth probabilities to continued increases in male income will decline. As in expression (3), this mixture of responses can be described for an aggregate population as

$$(4) \quad dB = K\frac{\partial B_2}{\partial Y_m}dY_m + (1-K)\frac{\partial B_1}{\partial Y_m}dY_m + \left[\alpha\frac{\partial B_2}{\partial Y_m} + (1-\alpha)\frac{\partial B_1}{\partial Y_m}\right]\frac{\partial K}{\partial Y_m}(dY_m)^2$$

where, as before, we write the movement ΔK as $(\partial K/\partial Y_m)dY_m$.

The last expressions in both (3) and (4) are proportional to squared differentials. By

ignoring these "small" terms and combining (3) and (4), we can write an aggregate birth probability equation in elasticity form as

$$(5)\quad dlnB = \left(\frac{B_2}{B}\eta_{B_2Y_m}\right) K \cdot dlnY_m + \left(\frac{B_1}{B}\eta_{B_1Y_m}\right)(1-K)\cdot dlnY_m + \left(\frac{B_2}{B}\eta_{B_2W_f}\right) K \cdot dlnW_f$$

Percentage changes in male income affect the average birth probability as a weighted average of the male income elasticities for couples where the wife is and is not employed. Percentage changes in women's wages affect the average birth probability in proportion to the employment ratio of women.

The dependent variable in equation (5) corresponds to the average probability of a birth. For estimation, we replace this measure with its population analogue, the fertility rate. Hence, at the aggregate level, equation (5) suggests the following interaction model for estimation purposes:

$$(6)\quad lnB = \beta_0 + \beta_1 K \cdot lnY_m + \beta_2(1-K)\cdot lnY_m + \beta_3 K \cdot lnW_f$$

where we hypothesize that $\beta_1 > 0$, $\beta_2 > 0$, $\beta_3 < 0$, and all data are in the form of natural logarithms.[8]

Collecting terms in $K \cdot lnY_m$, we obtain

$$(7)\quad lnB = \gamma_0 + \gamma_1 K \cdot lnY_m + \gamma_2 lnY_m + \gamma_3 K \cdot lnW_f$$

where $\gamma_1 = \beta_1 - \beta_2$, $\gamma_2 = \beta_2$, and $\gamma_3 = \beta_3$. An implication of this formulation is that $\gamma_1 + \gamma_2 > 0$.

The total response of the *population* fertility rate to changes in male income and female wages is given by

$$(8)\quad \eta_{BY_m} = \gamma_1 K + \gamma_2 = (\beta_1 - \beta_2)K + \beta_2$$

$$(9)\quad \eta_{BW_f} = \gamma_3 K = \beta_3 K$$

The critical feature of this formulation is that the effect of female wages on fertility rates is proportional to the female employment ratio. Consider then the effect of a cyclical movement in the labor market that raises the earnings of all participants. When few women are employed, annual changes in family income consist mainly of changes in the earnings of men. To the extent that the costly activity of childbirth is chosen to occur when family income is high, high-income periods are high-fertility periods. Alternatively, if a large proportion of the work force is women, annual changes in family income result significantly from changes in the earnings of women. Increases in women's earnings, like husbands' earnings, are expected to have a positive income effect on fertility. However, women's earnings also represent a cost of childbearing and rearing, for they come at the expense of time spent at home that could be used in child care. Hence, when female wages rise, families substitute against children and in favor of items that require less of the wife's time. Good times economically are the most expensive times to have children for women who are employed or on the margin of becoming employed. The larger the proportion of such women in the population, the greater the likelihood that good times will be associated with low-fertility rates for the whole population.

For the *U.S.* case, the remarkable postwar growth in employment of women of childbearing age may have been sufficient to weaken and then reverse the longstanding positive association between business cycles and fertility.

II. Determinants of Age-Specific Fertility Rates: 1947–74

Table 1 presents regression results and Table 2 presents population fertility rate elasticities derived from the specification of equa-

[8]The conjecture that β_3 is negative is simply an empirical proposition. An increase in the wife's wage has an income effect proportional to her hours worked in the market place, in addition to a pure price effect—the usual Slutzky decomposition. The sum of these two effects may be positive or negative. An obvious feature of this formulation is that the income effect may grow over time if women's hours are trending upward—for example, if working women move from part-time to full-time employment. We have ignored this element in our model.

TABLE 1—REGRESSIONS FOR AGE-SPECIFIC AND TOTAL FERTILITY RATES, 1948–75[a]

Independent Variables	20–24	25–34	35–39	Total Fertility Rate
$K \cdot lnW_f$	−1.759	−2.578	1.038	−4.745
	(−2.25)	(−2.14)	(0 27)	(−2.93)
$K \cdot lnY_m$	−.448	.104	−1.927	−.0239
	(−3.01)	(0.37)	(−1.62)	(−.04)
lnY_m	.876	.270	2.264	1.316
	(4.54)	(2.43)	(3.32)	(2.51)
Intercept	.0313	2.659	−8.84	−4.570
	(.02)	(3.79)	(−2.40)	(−1.64)
R^2	.93	.94	.88	.95
Durbin-Watson Statistic	0.74	1.00	0.67	1.53
Degrees of Freedom	24	24	24	24

[a]Asymptotic t-statistics in parentheses.

TABLE 2—POPULATION ELASTICITY ESTIMATES, 1948–75[a]

Elasticity of Fertility Rate with Respect to:	Age Group 20–24	25–34	35–39	Total Fertility Rate
Female Hourly Earnings				
1948–75	−820	−.956	.450	−1.732
	(−2.25)	(−2.14)	(0.27)	(−2.93)
1948–60	−.751	−.851	.409	−1.590
	(−2.25)	(−2.14)	(0.27)	(−2.93)
1960–75	−.871	−1 036	.482	−1.846
	(−2.25)	(−2.14)	(0.27)	(−2.93)
Male Annual Earnings				
1948–75	668	.308	1.428	1.308
	(4.03)	(3.43)	(3.32)	(4.24)
1948–60	.685	.304	1.505	1.308
	(4 10)	(3.52)	(3.47)	(4.03)
1960–75	.655	.312	1 370	1.307
	(3 98)	(3.34)	(3.18)	(4.37)

[a]See Table 1, All estimates are derived from those in Table 1.

tion (7).[9] The dependent variables are natural logs of fertility rates for female age groups, 20–24, 25–34, and 35–39. In addition, we present estimates for the total fertility rate.[10] We treat K, the proportion of couples with employed wives, as an endogenous variable and estimate all equations using two-stage least squares.[11] Since K is changing over time, the population elasticities are computed for two subperiods 1948–60 and 1960–75, in addition to the full sample, by using the average values of K within those periods. All wage and income variables are in constant 1964 dollars.

[9]A description of the sources and construction of the data is available from the authors on request. For the 20–24 age group, the male income variable is constructed as a weighted average of median income of males aged 20–24 and 25–34, with weights .57 and .43, respectively. These weights are the average probabilities for the 1950–70 period that a child born to a woman aged 20–24 has a father aged 25–34 at the time of the child's birth. For the oldest group, 35–39, the independent variables are available only for an overlapping group, 35–44. Independent variables in the total fertility rate regression are for ages 16 and over. While the dependent variables are constrained to lie between zero and 1,000, their range in data is between 20.2 and 358. The difference between estimates with a linear probability model and, for example, a logistic transformation of the data are extremely small.

[10]The total fertility rate is the sum of age-specific fertility rates. It represents the number of children a woman would have if she moved through her childbearing years experiencing the age-specific fertility rates in a particular calendar year. It would be interesting to investigate the determinants of marital fertility rates using explanatory variables for married couples. Unfortunately, such aggregate time-series data are not available by marital status.

[11]All results treat the products $K \cdot lnY_m$ and $K \cdot lnW_f$ as endogenous variables. The list of instrumental variables for the results in Table 1 are lnY_{m_t}, lnW_{f_t}, $lnY_{m_{t-1}}$, and $lnW_{f_{t-1}}$. The implicit model used to estimate K, the proportion of employed women, includes lagged values of women's wages and men's income. The justification for this rests on the theoretical observation that fixed costs of participating in the labor force would induce state dependence in sequential participation probabilities. In addition, there is some evidence (see James Heckman and Willis, 1977) from micro data, that this effect operates. The formally correct instrument list would then include current exogenous variables and lagged female employment. In fact, results using this instrument list are superior in the sense that standard errors are somewhat lower for all age groups and the results for the oldest age group, 35–39, reported below, conform with those of the younger groups. However, since the data on employment ratios are highly serially correlated, there remains some doubt that K has been "purged" of its endogenous component. We have chosen, therefore, to use additional instruments for lagged employment, i.e., lagged women's wages and lagged male income. An F-test for the appropriateness of excluding these lagged values from the fertility equation (see Robert Basmann) yields values for the age groups 20–24, 25–34, 35–39 of 0.11, 2.98, and 1.77, respectively. All of these are below the critical 5 percent level of $F(1, 23) = 4.28$.

Results for the two younger age groups and the total fertility rate conform with our hypotheses. The elasticity of fertility with respect to male income is significantly positive, while the female wage elasticities are negative, significant at the 5 percent level, and rising in magnitude over time. The regression results for the oldest group, 35–39, are less satisfactory. Though male income effects are still positive, female wage effects have "incorrect" signs and are insignificantly different from zero. Abstracting from the dynamic considerations mentioned above may particularly affect the results for this older group. Current economic variables may be far less important determinants of current fertility for couples whose fecund lifetime is nearing completion. In this case, the difference between total desired and current numbers of children, the stock effect, would become more important.

For this reason, we have focused our attention on the youngest age group, which contributes more than one-third of the total number of births. The upper panel of Figure 2 plots the data used to construct the explanatory variables in the 20–24 regression. The lower panel plots the actual 20–24 fertility rate and fitted values from this regression.

Although the predictive performance of the model for this age group is quite high, traditional goodness-of-fit measures are somewhat suspect in time-series models. Accordingly, we have estimated the same model in an early subperiod of the sample and predicted beyond the estimation period. The lower panel of

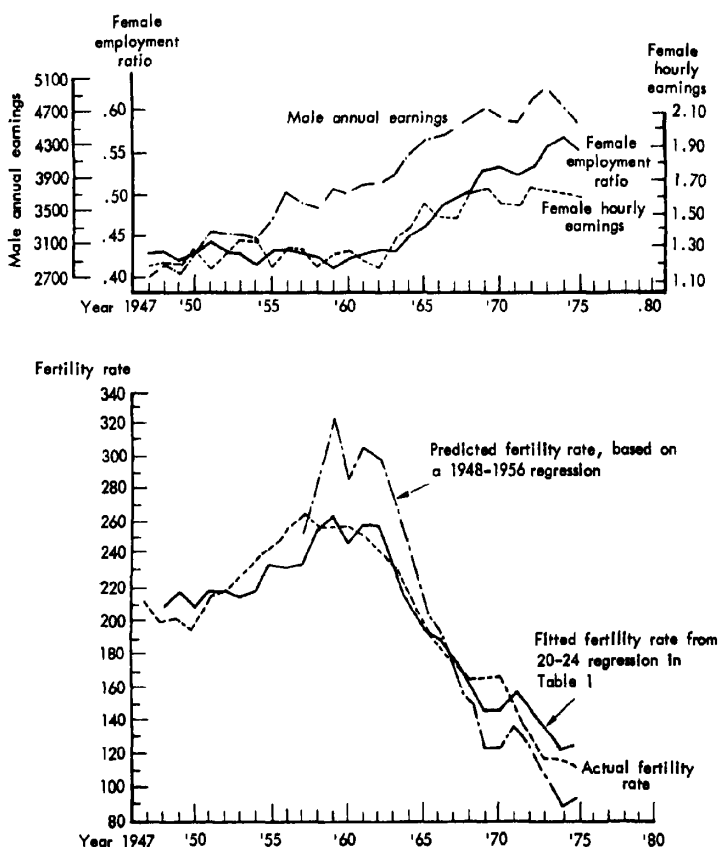



FIGURE 2. FERTILITY RATE, EMPLOYMENT RATIO, and HOURLY EARNINGS OF WOMEN AGED 20–24; ANNUAL EARNINGS OF THEIR HUSBANDS; AND FITTED AND PREDICTED FERTILITY RATES, 1947–75

Figure 2 plots these predicted rates for 1957–75 based on estimates of the same specification reported in Table 1, but using data only from 1948–56, one year before the peak of the baby boom.[12] Other than short lags at the late 1950's peak and the late 1960's and mid-1970's levelings, and an exaggeration of the peak, the equation fitted on these early data quite accurately traces the next nineteen years of fertility.[13] This predicted series simulates the fertility decline during the economic prosperity of the mid-1960's and the leveling in the late 1960's during a period of falling male earnings. Likewise, during the economic expansion following 1971, both predicted and actual fertility resumed their decline. The 1974–75 leveling of the actual series is matched by a predicted increase during a period of economic recession. *The conventional wisdom that period fertility moves procyclically would have predicted the opposite movements in each of these periods.*

While the results reported here assume that only contemporaneous values of the independent variables affect period fertility rates, we have also estimated specifications with one- or two-year lagged values of explanatory variables.[14] The contemporaneous specifications consistently track somewhat better. Previous studies calculating the correlations between detrended birth rates and single measures of aggregate economic activity have reported results with only one- or two-year lagged values of the latter variables.[15] However, our results, which stress the distinction between male and female earnings, suggest that fertility movements may be more closely associated with contemporaneous variations in the economic environment.[16]

III. Alternative Explanations of Postwar Fertility Movements

Although fertility rates and employment ratios of women are recognized to be strong negative correlates, there is much debate over the path of causation between these two variables. For example, it is often argued that the sharp increase in employment ratios of young women in the 1960's was a *result* of their declining fertility. The driving force behind the fertility decline is frequently taken to be improvements in contraceptive technology,[17] increased supply of modern contraceptives, or changes in women's tastes regarding children. The model we estimate takes the different view that women entered the labor market and reduced their fertility in the 1960's *because* of increasingly higher wages. The results discussed above support this position, and there is other indirect evidence.[18]

Table 3 shows correlations between fertility rates, female employment ratios, and female hourly earnings in both levels and first differences for 1947–75. If postwar movements in female employment were caused primarily by

[12]The regression equation predicts progressively better as the period of estimation lengthens. Going in the other direction, one can predict a 1960's fertility decline on as little basis as the experience of 1947–53, although the predicted decrease is then much less steep than what actually occurred.

[13]The squared correlation between actual fertility, 20–24, and the predicted series outside the range of estimation is .94. Henri Theil's inequality coefficient for these two series is only .016.

[14]We are prevented from developing a richer dynamic specification because of the short time-series and severe multicollinearity.

[15]See the studies cited in fn. 1. Only Thomas and Ogburn and Silver report that contemporaneous values of explanatory variables were inferior to the lagged formulations.

[16]We also attempted other variations on the first regression in Table 1. First, average schooling level of women age 20–24 was added as a determinant of birth probabilities, increasing the estimated income elasticity, but not altering the wage elasticity. Schooling itself has a negative effect, as the traditional hypotheses suggest, although it is insignificant. Almost no explanatory power is added to the basic model. Second, we estimated the model with different definitions of male income. Given the choice between median annual earnings of men age 20–24 and of men age 25–34, we ran regressions using each separately as well as the weighted average used in the regression in Table 1. In most specifications, the sum of squared residuals is minimized with a male income variable weighted more strongly toward the older age group than the variable we use. This may be because fewer married men than single men age 20–24 are part-time workers and students. If so, young married men's income would move like the income of the next older age group.

[17]See Norman Ryder (1973, 1974); Charles Westoff.

[18]Given the short time-series of data, we are prevented from applying a more sophisticated test for causal ordering. See Christopher Sims.

TABLE 3—CORRELATIONS OF FERTILITY RATES B, FEMALE EMPLOYMENT RATIOS K, AND FEMALE HOURLY EARNINGS W, 1947–75, IN LEVELS AND FIRST DIFFERENCES, BY AGE GROUP

Age Group	Levels of Variables			First Difference of Variables		
	$r(B,K)$	$r(B,W)$	$r(K,W)$	$r(\Delta B,\Delta K)$	$r(\Delta B,\Delta W)$	$r(\Delta K,\Delta W)$
20–24	−.92	−83	.92	−.28	−.33	−.01
25–34	−.93	−.94	.98	−.40	−33	.37
35–39	−.92	−.92	.97	−.18	−.01	.31
20–44	−82	−90	.98	−.07	−.21	.32

shifts in the supply function of female labor due to independent fertility changes caused by improved contraceptives or "taste" factors, then increases in the employment ratio would be accompanied by declines (or relative declines) in the female wage. Alternatively, employment changes generated by shifts in the demand for labor would cause wages to move in the same direction as the employment ratio but in the opposite direction to fertility, as long as fertility and female employment are negatively related.

Table 3 indicates that women's fertility has been strongly negatively correlated with their employment ratios and hourly earnings in the postwar period. In addition, the employment ratio and hourly earnings are positively correlated in levels and first differences. This supports the hypothesized predominance of labor demand shifts in producing female employment and fertility changes and is inconsistent with explanations of the baby bust that emphasize supply-side factors.[19]

In another explanation of postwar fertility movements, Richard Easterlin (1968, 1972) postulates that the fertility behavior of married couples is primarily determined by the relationship between their current and recent economic well-being and the well-being they expected. The thrust of his hypothesis is that young couples establish minimal expected standards of living on the basis of their experience as children growing up in their parents' household. If these expectations are not met, young couples will defer births, having children of their own only if this minimal income level is exceeded.[20]

We have used two proxies for relative economic status, constructed by Easterlin (1976) and Michael Wachter, in the regressions reported in Table 4.[21] Easterlin's measure is the difference between a twenty-year moving average (ending three years before the current date) of the general unemployment rate and an eight-year moving average of the rate. The second average measures young persons' standard of living and the first their desired or expected standard of living, as determined by unemployment conditions for their parents when these young people were still in their parents' home. Wachter proxied the relative status variable with a ratio of current expected hourly wage of all persons to

[19]Furthermore, the first major postwar breakthrough in contraceptive technology was the oral birth control pill. It was not authorized for public use until June 1960 and could not have had a noticeable effect on fertility for another eighteen months at the earliest. The proportion of married women under age 45 using it, as of October of the successive years from 1960 through 1965 was: 0.5 percent, 1.1 percent, 2.7 percent, 6.3 percent, 10.9 percent, and 17.9 percent (see Ryder and Westoff, 1967, p ii). Yet the postwar fertility leveling and decline began in 1958 (see Easterlin, 1972, p. 187; Ryder, 1969, p. 595; C. V. Kiser, Wilson Grabill and Arthur Campbell, p. 53). Easterlin (1972, pp. 187–91) provides an insightful analysis of the role of the oral contraceptive in the recent fertility decline.

[20]Of course, it could also be argued that young couples establish minimal desired family size on the basis of the number of their siblings, and that they only demand other "luxury" goods if their income rises above that necessary to sustain expected family size. See Yoram Ben-Porath for evidence that number of siblings influences desired fertility.

[21]Easterlin (1968, 1972) has also proposed two other more direct measures of the income of young couples relative to that of their parents. Unfortunately, these series cannot be constructed for the entire postwar period. In addition Ronald Lee and Peter Lindert have made important empirical contributions along these lines.

TABLE 4—ELASTICITY ESTIMATES INCLUDING PROXIES FOR THE RELATIVE ECONOMIC STATUS VARIABLE, 1948–74
(Asymptotic t-statistics in parentheses)

Elasticity of Fertility Rate 20–24 with respect to:	Regression Numbers				
	(1)	(2)	(3)	(4)	(5)
Male Annual Earnings	.634	1.00		.691	
	(3 61)	(4 74)		(4 14)	
Female Hourly Earnings	−.728	− 917	−1 131	−.736	−1.07
	(−2.17)	(−2.69)	(−5.21)	(−2.32)	(−6.12)
Easterlin's Unemployment Difference Variable		.042	.010		
		(2 22)	(.428)		
Wachter's Relative Wage Variable				1.306	1.271
				(2 34)	(1.28)
R^2 of the Regression	92	.94	.74	.93	.76
Durbin-Watson Statistic	67	1.09	.33	.81	.33
Degrees of Freedom	23	22	24	22	24

a ten-year moving average (ending in current year) of the expected wage. His expected wage is corrected for the unemployment rate.[22]

Regression (1) repeats our basic regression (age 20–24) from Table 1.[23] The other regressions include the proxies for relative status as determinants of birth probabilities of employed and nonemployed women. Comparing regression (2) with regression (1) shows that Easterlin's unemployment difference variable has a significant positive coefficient as hypothesized and strengthens the estimated male income and female wage elasticities. However, removing all current male income variables in regression (3) reduces the unemployment difference to insignificance but strengthens the coefficient of female wage.

Regressions (4) and (5) repeat these tests with the Wachter relative wage variable. Regression (4) shows a significant positive elasticity with respect to this variable, as well as stable estimates of the income and wage elasticities. However, the Wachter variable does less well when the male earnings variables are excluded in regression (5).

The results in Table 4 strongly indicate that both current income and female wage operate independently of relative status variables. The latter also play a role, though not a dominant or completely consistent one.

V. Conclusions and Implications

Two conclusions can be drawn from this study. First, economic models of fertility behavior that emphasize the distinction between income and price effects—what we have called male income and female wage effects—are successful in explaining not only important features of cross-sectional variations but also aggregate movements of *U.S.* fertility over time. Second, the model we propose successfully predicts both procyclical and countercyclical variations in fertility in a unified framework, across age groups, and over a long time span.

The current level of employment of young

[22]Wachter tests the importance of a relative status variable in a more complete economic model of time-series fertility. His principal variables are his relative wage variable (described above), the percent of the population living in urban areas, interactions among these and their squares, and the predicted dependent variable (fertility rate) lagged one year. Wachter's estimated equations track fertility very well from 1927–72. However, most of the equations' tracking ability appears due to the lagged predicted dependent variable. In addition, interpretation of the urbanization variable is difficult, as Wachter points out. At various points, he interprets it as a proxy for the relative price of children, for parental taste for children, and for the contraception failure rate.

[23]The sample for Table 4 excludes 1975 since we do not have values of the relative status variables for that year.

women and variation in their wages are more than high enough to induce continuing countercyclical fertility movements. The fertility increases or levelings of 1970 and 1974 reflect this phenomenon, and we see no reason to expect a change so long as a large proportion of young women are employed. In addition, we expect the female employment ratio to continue its secular increase and female wages to rise as long as the economy expands. Predictions for the more distant future must account for a potential bottoming out of the secular decline in fertility. Certainly these rates will not reach zero. However, we expect a continuing secular decline in fertility toward this asymptote, punctuated by countercyclical fertility movements.

This view differs from that of June Sklar and Beth Berkov, who interpret the 1974 fertility leveling as evidence of a coming secular rise, as successive cohorts of women 30–34 find themselves childless with little time left to begin a family. Our expectations are also the opposite of Wachter, whose work implies that "a new upswing in the rate of growth in real wages could cause a new 'baby boom' as relative wages increase and offset the drag on fertility of the increasing cost of children" (p. 623) (represented by urbanization). Similarly, Easterlin wrote: "It seems unlikely that over the next decade age-specific birth rates of young persons will decline much below the level at the beginning of the decade. If a sustained economic boom were to take place . . . a rise in these rates is possible" (1972, pp. 212, 216). More recently, however, Easterlin (1976, pp. 419–20) has predicted a continued fertility decline through the late 1970's. Finally, Lee projects, based on the relative income hypothesis, that "current U.S. fertility is only temporarily low, and an upturn should occur after five or ten years" (p. 467).

In our view, the only changes that would produce the leveling or increasing fertility that these authors expect are a large fall in young women's employment, which would restore the traditional positive relationship between aggregate economic conditions and fertility, or a substantial increase in the supply of preschool or day care facilities, which might weaken the link between women's market wages and the price of children. The first is very unlikely. The second may be happening over time,[24] although available evidence suggests that parents in the United States continue to place high value on bearing and rearing their own children and continue to respond to changes in the costs of this activity.

[24]For example, the proportion of children aged 3 and 4 enrolled in school in October increased threefold between 1964 and 1974, standing at almost 30 percent in the later year.

REFERENCES

R. L. Basmann, "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," *J. Amer. Statist. Assn.*, Dec. 1960, *55*, 650–59.

G. S. Becker, "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, Universities–Nat. Bur. Econ. Res. conference series 11, Princeton 1960, 209–31.

Y. Ben-Porath, "First Generation Effects on Second Generation Fertility," *Demography*, Aug. 1975, *12*, 397–405.

Richard A. Easterlin, "Relative Economic Status and the American Fertility Swing," in Eleanor B. Sheldon, ed., *Social Structure, Family Life Styles, and Economic Behavior*, Philadelphia 1972.

———, "The Conflict Between Aspirations and Resources," *Pop. Develop. Rev.*, Sept./Dec. 1976, *2*, 417–26.

———, *Population, Labor Force, and Long Swings in Economic Growth: The American Experience*, New York 1968.

V. L. Galbraith and D. S. Thomas, "Birth Rates and Inter-War Business Cycles," *J. Amer. Statist. Assn.*, Dec. 1941, *36*, 465–76.

J. J. Heckman and R. J. Willis, "A Beta-logistic Model for the Analysis of Sequential Labor Force Participation by Married Women," *J. Polit. Econ.*, Feb. 1977, *85*, 27–58.

H. Hyrenius, "The Relation Between Birth Rates and Economic Activity in Sweden 1920–1944," *Bull. Oxford Univ. Instit. Econ. Statist.*, Jan. 1946, *8*, 14–21.

D. Kirk, "The Relation of Employment Levels to Births in Germany," *Milbank Memorial*

Fund Quart., Apr. 1942, *20*, 126–30; reprinted in Joseph A. Spengler and Otis D. Duncan, eds., *Demographic Analysis: Selected Readings*, Glencoe 1963, 195–206.

———, "The Influence of Business Cycles on Marriage and Birth Rates," in *Demographic and Economic Change in Developed Countries*, Princeton 1960, 214–57.

C. V. Kiser, Wilson H. Grabill, and Arthur A. Campbell, *Trends and Variations in Fertility in the United States*, Cambridge, Mass. 1968.

R. D. Lee, "Demographic Forecasting and the Easterlin Hypothesis," *Pop. Develop. Rev.*, Sept./Dec. 1976, *2*, 459–68.

Peter H. Lindert, *Fertility and Scarcity in America*, Princeton 1977.

J. Mincer, "Market Prices, Opportunity Costs, and Income Effects," in Carl Christ et al., eds., *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grunfeld*, Stanford 1963.

Norman Ryder, "The Emergence of a Modern Fertility Pattern: United States 1917–1966," in S. J. Behrman, Leslie Corsa, and Ronald Freedman, eds., *Fertility and Family Planning: A World View*, Ann Arbor 1969, 99–117.

———, "Recent Trends and Group Differences in Fertility," in Charles F. Westoff, ed., *Towards the End of Growth*, Englewood Cliffs 1973, 57–68.

———, "The Family in Developed Countries," *Scientific Amer.*, Sept. 1974, *231*, 123–32.

——— **and Charles F. Westoff,** "The United States: The Pill and the Birth Rate, 1960–1965," *Stud. Family Planning*, No. 20, June 1967.

J. Sklar and B. Berkov, "The American Birth Rate: Evidence of a Coming Rise," *Science*, Aug. 29, 1975, *189*, 693–700.

T. P. Schultz, "Fertility Determinants: A Theory, Evidence and an Application to Policy Evaluation," R-1016-RF/AID, Rand Corp., Jan. 1974.

Theodore W. Schultz, *Economics of the Family*, Chicago 1974.

M. Silver, "Births, Marriages, and Business Cycles in the United States," *J. Polit, Econ.*, June 1965, *73*, 237–55.

C. A. Sims, "Money, Income, and Causality," *Amer. Econ. Rev.*, Sept. 1972, *62*, 540–52.

A. Sweezy, "The Economic Explanation of Fertility Changes in the United States," *Pop. Stud.*, July 1971, *25*, 255–67.

Henri Theil, *Applied Economic Forecasting*, Chicago 1966.

Dorothy S. Thomas, *Social Aspects of the Business Cycle*, New York 1927.

——— **and W. F. Ogburn,** "The Influence of the Business Cycle on Certain Social Conditions," *J. Amer. Statist. Assn.*, Sept. 1922, *66*, 324–40.

M. L. Wachter, "A Time-Series Fertility Equation: The Potential for a Baby-Boom in the 1980's," *Int. Econ. Rev.*, Oct. 1975, *16*, 609–24.

M. P. Ward and W. P. Butz, "Completed Fertility and its Timing: An Economic Analysis of U.S. Experience Since World War II," R-2285-NICHD, Rand Corp., Apr. 1978.

C. F. Westoff, "The Decline of Unplanned Births in the United States," *Science*, Jan. 9, 1976, *191*, 38–41.

R. J. Willis, "A New Approach to the Economic Theory of Fertility Behavior," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, Part II, 514–64.

———, "The Economic Determinants of Fertility Behavior," unpublished doctoral dissertation, Univ. Washington 1971.

The Rate of Surplus Value, the Organic Composition, and the General Rate of Profit in the U.S. Economy, 1947–67

By EDWARD N. WOLFF*

Marx's "law of the tendency of the rate of profit to fall" has been the subject of much debate, both theoretical and empirical. The law states that over time the organic composition of capital will tend to rise, thereby causing the general rate of profit to fall (vol. 3, ch. 13). At issue on the theoretical side are the analytical relations between the organic composition of capital and the general rate of profit and, more broadly, between labor values and prices of production. At issue on the empirical side are the actual movements of the organic composition and the general rate of profit over time. This paper will investigate both sides of this issue.

The first part will develop and criticize the law of the falling rate of profit. The presentation will rely heavily on the analytical apparatus developed for Marxian analysis over the last twenty years (see particularly Francis Seton, Michio Morishima and Seton, Paul Samuelson, Morishima, William Baumol, and John Roemer), as well as Jens Christiansen's review of the arguments. The major conclusion that will be drawn is that the general rate of profit does not necessarily move inversely to the organic composition of capital and there is thus no necessity for the rate of profit to fall with capitalist development.

What remains of the "law" of the falling rate of profit is then the empirical question of whether in fact the rate of profit had fallen and, if so, what factors have contributed to its decline.[1] The second part of the paper will present empirical estimates of the general rate of profit, the organic composition of capital, and other key variables in the Marxian system for the 1947–67 period in the United States. The study is confined to this period because the required data (input-output tables) are available only for years 1947, 1958, 1963, and 1967. Two questions will be addressed. First, has the rate of profit fallen? Second, what is the observed relation of the rate of profit to the organic composition and other variables in the system? In this regard, particular attention will be paid to relative productivity increases over the period and their impact on the rate of profit, the organic composition, the rate of surplus value and other variables.

I. The Law of the Falling Rate of Profit

A. *A Two-Sector Model*

For heuristic reasons, the law of the falling rate of profit will be analyzed in a simple two-sector economy. Suppose that the first sector produces wage (consumer) goods and the second sector produces capital goods; that capital goods depreciate at the rate δ and are used in both sectors and there are no other interindustry flows; and that labor is homo-

*Assistant professor, New York University. An early draft of this paper was presented at the political economy seminar at Stanford University. I would like to thank those present, as well as William Baumol, Wassily Leontief, George Borts, and an anonymous referee for their helpful comments and suggestions. I would also like to thank Anne Carter and Peter Petri for supplying the data used in this study.

[1]The debate over the secular tendency of the rate of profit is not confined to Marxian analysts. William Nordhaus has argued that there is a secular tendency for the rate of profit to fall, and this is observable for the *U.S.* economy in the 1966–73 period. Martin Feldstein and Lawrence Summers, with more recent data and a different set of adjustments, have countered that there is no long-run tendency for the rate of profit to fall, only a cyclical variation in the profit rate due to business cycle forces. The major period of contention, the post-1966 period, unfortunately, falls outside the scope of this paper.

geneous and is used in both industries. Thus, the interindustry coefficient matrix a is given by

$$a = \begin{pmatrix} 0 & 0 \\ \delta k_1 & \delta k_2 \end{pmatrix}$$

where k_i (the capital coefficient) is the amount of capital required per unit of output in sector i. The vector of labor coefficients l is given by $l = (l_1 l_2)$ where l_i is the amount of labor required per unit of output in sector i.

Marx's law of the falling rate of profit is formulated in terms of two sets of relative prices. The first is the set of labor values ("values" for short) and the second is the set of prices of production ("prices" for short). The key variable of interest, the "general rate of profit," is a price of production variable, and this set of prices will be developed first.

Marx's prices of production ρ are the vector of relative prices such that the rate of profit on the capital advanced in each sector is equal. The rate of profit that satisfies this condition is the general rate of profit π. Thus ρ is the set of prices and π the rate of profit that would be observed if there were perfect competition in the economy. For simplicity, it will be assumed for now that this rate of profit is computed on current costs (see Section IB). Let ρ_1 be unity and assume that the real wage w (which is also the money wage in this case) is exogenous. Then, ρ and π are given by

$$(1) \qquad (1 + \pi)(wl_1 + \delta\rho_2 k_1) = 1$$

$$(2) \qquad (1 + \pi)(wl_2 + \delta\rho_2 k_2) = \rho_2$$

The solution is implicit from the eigen equation:

$$(3) \quad \rho B = s\rho$$

where $s = 1/(1 + \pi)$

$$B = \begin{pmatrix} wl_1 & wl_2 \\ \delta k_1 & \delta k_2 \end{pmatrix}$$

In the two-sector case, the solution for s is analytic and is given by

$$(4) \quad 2s = wl_1 + \delta k_2 \pm [(wl_1 - \delta k_2)^2 + 4(wl_2\delta k_1)]^{1/2}$$

From equation (1),

$$\rho_2 = \frac{s - wl_1}{\delta k_1}$$

The condition that $\rho_2 > 0$ implies that

$$(\delta k_2 - wl_1) + [(\delta k_2 - wl_1)^2 + 4(wl_2\delta k_1)]^{1/2} > 0$$

Thus, only the positive root in equation (4) satisfies the condition that ρ_2 is positive. The solution to this special two-sector case demonstrates that positive prices and a positive profit rate not only exist, but are unique.[2]

The other major variable in Marx's law of the falling rate of profit is the organic composition of capital. This variable is derived from Marx's labor value framework. The labor value of a commodity, also referred to as its labor content, its total labor requirement, and the labor embodied in it, is defined as the direct plus indirect labor required to produce it. These values are never observed, since their formation violates the condition of an equal rate of profit in all sectors. They are, however, of vital significance in the Marxian system, as will be discussed below.

The labor value of commodity 2, λ_2, is given by

$$\lambda_2 = l_2 + \lambda_2\delta k_2$$

$$\lambda_2 = \frac{l_2}{1 - \delta k_2}$$

The labor value of commodity 1, λ_1, is given by

$$\lambda_1 = l_1 + \lambda_2\delta k_1$$

$$\lambda_1 = l_1 + \frac{l_2\delta k_1}{1 - \delta k_2}$$

The value of labor power is defined as the cost of reproducing labor or, equivalently, as the labor content of worker consumption. In this case, it is given by $\lambda_1 w$. One of Marx's basic insights about the capitalist system is that the amount of time actually worked by labor is greater than the labor time equivalent of their

[2]In the general n-sector case, both the existence and the uniqueness of a set of positive prices of production and a positive profit rate can be demonstrated (see Morishima and Seton). The solution is given by the dominant (Frobenius) eigenvalue of equation (3) (see Akira Takayama, pp. 367–79).

wage goods. This difference is the surplus value (labor time) S generated by labor and is given by $1 - \lambda_1 w$ per unit of time. The value advanced for the worker's consumption is called variable capital, V. The ratio of surplus value to variable capital is called the rate of surplus value ϵ and is given by

$$\epsilon = \frac{S}{V} = \frac{1 - \lambda_1 w}{\lambda_1 w} \tag{5}$$

In the case of homogeneous labor, the rate of surplus value is the same for all workers since the time worked and the real wage is the same.

Constant capital C is defined as the labor value of the fixed capital plus circulating capital (interindustry flows) used in production. The organic composition σ is defined as the ratio of the total constant capital to the total variable capital advanced in the economy per turnover period.[3] In this special case where there are no interindustry flows except depreciation, it is given by

$$\sigma = \frac{C}{V} = \frac{\lambda_2(kX)}{(\lambda_1 w)N}$$

Where $k = (k_1 k_2)$, $X' = (X_1 X_2)$ is the vector of total output by sector, and N is total employment. A related concept is the technical composition of capital τ, which is defined as the ratio of the "quantity" of physical capital to the amount of labor used in production. In this special case,

$$\tau = \frac{kX}{N}$$

The organic composition can thus be equivalently computed as

$$\sigma = \tau\left(\frac{\lambda_2}{\lambda_1 w}\right) \tag{6}$$

or

$$\sigma = (1 + \epsilon)\tau\lambda_2 \tag{7}$$

One final concept needs to be introduced; the value rate of profit π_v. This is defined as the ratio of total surplus value to the total capital advanced, both constant and variable. Dividing both numerator and denominator by V, we obtain

$$\pi_v = \frac{S}{C + V} = \frac{\epsilon}{\sigma + 1} \tag{8}$$

We are now in a position to understand Marx's law of the falling rate of profit. The law states that the organic composition of capital tends to rise over time, and thereby causing the general rate of profit (π) to fall. There are two points at issue in the law. The first is why the organic composition should tend to rise over time. There is good reason to suppose that the *technical* composition of capital will rise, since new technology tends to be labor saving. However, from equation (6) it is apparent that the organic composition will not rise if the real wage rises as fast as the technical composition, and the fall in labor values from increased productivity is of the same magnitude in the two sectors.

The second is why a rise in the organic composition would cause a fall in the general rate of profit. This was based on two beliefs. First, Marx believed that when labor values were "transformed" into prices of production, total surplus value, total variable capital, and total constant capital remained invariant. Thus, the general rate of profit π would equal the value rate of profit π_v. This invariance property, however, holds only under very restrictive conditions, as has been shown on numerous occasions.[4] Second, Marx believed that one could speak of a change in the organic composition independent of a change in the rate of surplus value. From equation (7), it is obvious that the two variables are not only functionally related but positively related. Thus, from equation (8), one cannot even show that a rise in the organic composition will cause a fall in the *value* rate of profit.[5]

[3]The turnover period is defined as the time between the start of the production period and the point at which the commodities are sold. It will be assumed here and throughout the paper that it is one year in each sector.

[4]See Samuelson, for example. Moreover, it should be noted that the value rate of profit is a function of the output mix (X), whereas the general rate of profit is independent of X. A shift in X would thus change the former but not the latter.

[5]See Morishima and Roemer for related discussions of the law of the falling rate of profit. Roemer does

In fact, it is quite easy to construct a case where the organic composition changes but the general rate of profit π remains unchanged. Suppose that the real wage w and technology (as defined by l, k, and δ) remain fixed but the *composition* of output shifts toward sector 2 (as, for example, from an autonomous increase in investment demand). If X^* is the new vector of output, then $X_2^*/X_1^* > X_2/X_1$. From equation (4) it is apparent that the general rate of profit will be unchanged, since it depends exclusively on the real wage and the technical coefficients. From equation (6) it can be seen that the organic composition will change proportionately to the technical composition, since the real wage is fixed and the labor values depend only on the technical coefficients. The technical composition τ^* becomes

$$\tau^* = \frac{kX^*}{lX^*}$$

and the organic composition σ^* becomes

$$\sigma^* = \frac{kX^*}{kX} \cdot \frac{lX}{lX^*} \cdot \sigma$$

Thus, as long as k_1/l_1 is different from k_2/l_2 (and, in general, this will be the case), the organic composition will change. In fact, if k_2/l_2 is greater than k_1/l_1, the organic composition will rise without π changing.

From this discussion it is apparent that Marx's law of the falling rate of profit cannot be supported on theoretical grounds. However, the possibility remains that the relations expressed in the law hold empirically. The concept of a labor value does embody the basic notion of worker productivity and the concept of the rate of surplus value does establish a basic relation between time worked and the compensation for that work.[6] Moreover, the concept of the organic composition does express a relation between the labor employed and the total capital advanced in production. If the "deviation" of prices of production from labor values is small, or large but not systematically biased with respect to the components of gross output, or even systematically biased but biased in the same direction over time, then trends in the organic composition should be directly related to trends in the general rate of profit. This is the subject of my next investigation. However, before proceeding, I must generalize the model to the n-sector case.

B. *Generalization to n Sectors*

The data used for the empirical analysis are standard Bureau of Economic Analysis (*BEA*) eighty-seven-sector *U.S.* input-output tables for years 1947, 1958, 1963, and 1967.[7] Since the interindustry data are in dollar terms, instead of physical terms, a Leontief input-output framework with several modifications as indicated below is used. Define

a = interindustry coefficient matrix, where depreciation is added as an endogenous row counterbalanced by an endogenous "depreciation replacement" column and an endogenous export column is added to balance the import row.[8]

l = row vector of labor coefficients.[9]

N = total employment.

m = column vector of average consump-

demonstrate that the general rate of profit is bounded by $\epsilon/(1+\sigma_{max})$ and $\epsilon/(1+\sigma_{min})$, where σ_{max} and σ_{min} are, respectively, the maximum and minimum sectoral organic compositions. Empirically, the spread in sectoral organic compositions is usually so great that their movement over time has very little bearing on the general rate of profit.

[6]In fact, this is one of the reasons why Marx begins his analysis of the capitalist system (*Capital*, vol. 1) with labor values and not with prices of production. Marx felt that basic social, institutional, and technological changes could be directly understood in terms of their impact on labor values and the rate of surplus value, not on the general rate of profit.

[7]All data, unless otherwise indicated, were obtained from Carter and Petri. Further details in data sources and adjustments may be found in my 1977b paper, or obtained from me.

[8]In the Marxian scheme, depreciation is treated as a produced input rather than as a part of value-added. Depreciation coefficients were obtained in Stephen Dresch and Robert Goldberg, Table A-1, and balanced using National Income and Product Account (*NIPA*) capital consumption allowance estimates. For details on the import adjustment see my 1975 paper. Moreover, I am assuming that all activity is productive (see my 1977b paper).

[9]Estimates for 1967 were obtained in the U.S. Bureau of Labor Statistics.

tion per worker. The total worker consumption vector $M = mN$.[10]

Y = column vector of surplus final demand (total final demand less total worker consumption M); total surplus income $\Pi = \Sigma Y$.

X = column vector of gross domestic output by sector.

k = capital coefficient matrix.[11]

p = row vector of sectoral price indices, with base year 1958.

Then the vector of labor values λ is given by

$$(9) \qquad \lambda = l(I - a)^{-1}$$

where λ_i is interpreted as the direct plus indirect labor required per dollar of output (in current prices) in sector i.[12] The value of labor power, or the variable capital advanced per worker, is equal to λm. The rate of surplus value ϵ is then given by

$$(10) \qquad \epsilon = \frac{1 - \lambda m}{\lambda m}$$

which may then be viewed as the ratio of uncompensated (surplus) to compensated (necessary) labor time.[13] Total variable capital, V, and total surplus value, S, are given by

$$(11) \qquad V = N \lambda m$$

$$(12) \qquad S = N \lambda m \epsilon$$

where N is total employment.[14]

The organic or value composition of capital σ is given by

$$(13) \qquad \sigma = \frac{\lambda(k + a)X}{N\lambda m}$$

where X is the column vector of total output by sector. The technical composition of capital τ is given by

$$(14) \qquad \tau = \frac{p(k + a)X}{N}$$

where constant prices have been used to measure the quantity of output. The relation between the two is given as

$$(15) \qquad \sigma = \frac{\lambda_c^*}{\lambda m} \tau$$

where λ_c^* is the average labor content of a dollar of physical capital in constant prices.[15] Alternatively, from (10):

$$(16) \qquad \sigma = (1 + \epsilon)\lambda_c^* \tau$$

Finally, the value rate of profit π_v is given by

$$(17) \qquad \pi_v = \frac{\epsilon}{\sigma + 1}$$

[10]Technically, M is the consumption by workers out of labor (wage and salary) earnings. To estimate this column, we computed total net wage and salary earnings (that is, net of personal income taxes) as a percent of total net personal income and applied this percentage to the household consumption vector of final demand. Data for this were obtained from *NIPA*.

[11]Years 1947 and 1958 were obtained from Carter and Petri. Years 1963 and 1967 were estimated by adjusting the 1958 coefficients for sectoral price changes and the change in the overall capital-output ratio (see John Kendrick, Table B-22).

[12]In my 1975 paper, following Morishima and Seton, I used a different procedure to estimate labor values in the case of the Puerto Rican economy, which was as follows: Let q be the matrix of interindustry *row* coefficients. Define matrix b such that

$$b_{ij} = \frac{Nm_i}{X_i} \frac{W_j}{\sum_j W_j}$$

where W_j is the total amount of net wages generated in sector j. The rate of surplus value ϵ and the vector of labor values λ is then given by

$$\left[\frac{1}{1+\epsilon} I - (1 - q')^{-1} b'\right]\lambda = 0$$

This method has the advantage over equation (9) of minimizing the distortions in computing λ that result from changing price structures. However, it has the disadvantage of being more sensitive to changing industrial wage structures. Some calculations will be made using this alternative technique and compared with those from equation (9).

[13]I have interpreted the value of labor power to refer to the actual labor earnings of workers, instead of the "subsistence" level of earnings, as Marx is sometimes construed. My interpretation maintains the essential relation between the real wage and the level of surplus at the disposition of the capitalist class, as well as avoiding difficult problems in measuring a subsistence level of consumption.

[14]It can easily be shown that $N = V + S$, the total value-added.

[15]That is,

$$\lambda_c^* = \frac{\lambda(k + a)X}{p(k + a)X}$$

TABLE 1—LABOR VALUE AND RELATED STATISTICS

	1947	1958	1963	1967
Rate of Surplus Value (ϵ)	1.009	1.048	1.102	1.122
Surplus/Worker Consumption ($\Sigma Y/\Sigma M$)	0.932	0.928	0.989	0.978
Organic Composition (σ)	5.52	5.85	5.67	5.78
Technical Composition (τ)	15.78	20.35	23.00	25.35
Average Labor Content of Constant Capital in Constant Prices (λ_c^*)[a]	0.174	0.140	0.117	0.107
Mean Earnings per Worker (thousands of 1958 dollars)	2.50	3.36	3.83	4.39
Value Rate of Profit (π_v)	0.155	0.153	0.165	0.166

[a]Statistics are in man-years per thousand dollars in 1958 prices.

Prices of production ρ and the general rate of profit π can be directly solved from the interindustry, capital, and labor coefficients and the real wage, as follows:

$$(\rho a + \rho k + \omega l)\,(1 + \pi) = \rho$$

where ω is the money wage in price of production terms. However, $\omega = \rho m$. Therefore,

$$\rho(a + k + ml) = \left(\frac{1}{1 + \pi}\right)\rho \tag{18}$$

Here, it should be made clear that the rate of profit is computed on the total capital advanced—that is, the sum of circulating capital (interindustry purchases), the net capital stock, and wages. Prices of production are thus equilibrium prices under the condition that the rate of the profit is equalized in each sector.[16]

II. Empirical Examination

A. *Labor Value Estimates*

Table 1 provides estimates of the rate of surplus value, the organic composition, and related variables for each of the four years. The rate of surplus value was approximately 100 percent during the period.[17] It rose 11 percent and at a relatively constant rate over the 1947–67 period. The ratio of surplus to worker consumption, on the other hand, rose only 5 percent over the period and at a nonuniform rate. It was uniformly lower than the rate of surplus value, reflecting the fact that the average labor content of labor's consumption was less than that of surplus consumption.[18]

The organic composition rose 5 percent at a nonuniform rate over the period, while the technical composition rose 61 percent and at an approximately uniform rate. The discrepancy in the two movements can be explained equivalently by the fall in the average labor content of constant capital, with a stable rate

[16]One can also formulate the problem, as Marx does, as a "transformation" of labor values into prices of production (vol. 3, ch. 9). The transformation then becomes

$$(\rho\lambda)(a + k + ml) = \left(\frac{1}{1 + \pi}\right)(\rho\lambda)$$

Furthermore, one can, following Marx, use an iterative algorithm to solve for ρ (see Morishima, ch. 7 for details, though note that he uses technical coefficients instead of dollar coefficients and does not include fixed capital).

[17]Estimates of the rate of surplus value made for other countries are of a similar magnitude: 0.93 for Japan in 1951 (see Nubua Okishio), 1.35 for Czechoslovakia in 1962 (See O. Kýn, B. Serkerka and L. Hejl), and 0.97 and 0.93 for Puerto Rico in 1948 and 1963, respectively, (see the author). This is not too surprising, given the simplifying assumptions used to construct the modified Leontief framework and given the broad equality of factor shares over a wide range of developed countries (see, for example, Jacques Lecaillon and Dimitrios Germidis). Moreover, estimates of the rate of surplus value, made following the alternative procedure outlined in fn 7, were similar to those in Table 1: 1.08, 1.16, 1.25, and 1.25 in each of the four years, respectively.

[18]The rate of surplus value can be equivalently formulated as $\epsilon = \lambda Y/\lambda M$.

TABLE 2—THE GENERAL RATE OF PROFIT AND OTHER MEASURES OF THE RATE OF PROFIT

	1947	1958	1963	1967
General Rate of Profit (π)	.139	.128	.137	.138
Value Rate of Profit (π_v)	.155	.153	.165	.166
Average Market Rate of Profit ($\Sigma\Pi/(\Sigma A + \Sigma K + \Sigma W)$)	.144	.132	.143	.145
Ratio of Surplus Income to Capital Stock ($\Sigma\Pi/\Sigma K$)	.290	.256	.283	.286
Ratio of Property Income to Capital Stock[a]	.211	.156	.150	.149
Ratio of Net Property Income to Capital Stock[b]	.163	.117	.106	.107
Ratio of After-Tax Corporate Profits to Capital Stock[c]	.055	.029	.035	.037
Nordhaus' "Genuine Rate of Return" on Nonfinancial Corporate Capital (after taxes)[d]	.097	.054	.081	.088
Feldstein and Summers' Net Rate of Return on Nonfinancial Corporate Capital[e]	.133	.085	.119	.199

[a]Property income is defined as the sum of business and professional income, farm income, rental income of persons, corporate profits and net interest before taxes. *Source*: U.S. Council of Economic Advisers, Table B-12, p 209

[b]Net property income is defined as property income less corporate taxes and personal income taxes paid on property income (See Appendix for estimation of taxes.) *Source*. U.S. Council of Economic Advisors, Table B-12, p. 209; Table B-15, p 212; Table B-17, p. 214; Table B-73, p. 280.

[c]After-tax profits include dividend payments *Source* U.S. Council of Economic Advisors, Table B-73, p. 280.

[d]The 1947 figure is actually for year 1948. *Source*: Nordhaus, p. 180

[e]The 1947 figure is actually for year 1948 *Source*. Feldstein and Summers, p. 216

of surplus value (equation (16)), or by the rise in the real wage (equation (15)).[19] The movement of the value rate of profit is a direct consequence of that of the rate of surplus value and the organic composition. It rose by 7 percent over the period, with the major increase occurring between 1958 and 1963.

B. *Prices of Production and the General Rate of Profits*

Prices of production were computed according to equation (18). One issue mentioned above is the deviation of relative prices of production from relative labor values. A regression of ρ_i on λ_i yielded R^2 of 0.97, 0.93, 0.92, and 0.91 in 1947, 1958, 1963, and 1967, respectively. This indicates that empirically relative prices were close to relative labor values for the *U.S.* economy during this period and that the general rate of profit as a result should be closely related to the value rate of profit.

This is confirmed in part (see Table 2). The general rate of profit declined by 1 percent over the 1947–67 period, while the value rate of profit increased by 7 percent. The general rate of profit was uniformly lower than the value rate of profit. However, both the general and value rates of profit fell between 1947 and 1958, rose between 1958 and 1963, and then increased slightly between 1963 and 1967.

[19]In the case of Puerto Rico, a rising technical composition and a falling organic composition were found, though only circulating capital was included in the calculations (see my 1977a paper). A comparison of Puerto Rico and the United States (with circulating capital only) shows that the organic composition in the two countries was relatively close, 2.8 and 2.1 in Puerto Rico in 1948 and 1963, respectively, and 2.3 and 2.3 in the United States in 1947 and 1963, respectively.The technical composition, on the other hand, was considerably higher in the United States, with values of 6.0 and 9.5 in 1947 and 1963, respectively. In Puerto Rico, the comparable figures were 2.0 in 1948 and 4.2 in 1963.

Comparisons were made with other concepts of the rate of profit. The first is the average market rate of profit, defined as the ratio of surplus income to the sum of circulating capital, capital stock, and wages in market price terms (line 3). Its value lay between the general and value rates of profit, and its movement paralleled that of the general rate of profit, with an 8 percent decline between 1947 and 1958 and a net change of less than 1 percent over the entire period. The second is the ratio of surplus income to the total capital stock (line 4). The difference between this measure and the average market rate of profit is that the denominator includes only fixed capital stock, which is closer in concept to more conventional measures of the rate of profit.[20] This ratio fell by 12 percent between 1947 and 1958 and then rose by 12 percent between 1958 and 1967, thus paralleling in movement the general rate of profit. The third concept is the ratio of before-tax property income, including corporate profits to total capital stock (line 5). In this measure, personal income taxes paid on wages and salaries are not included in the numerator. This ratio declined by 12 percent between 1947 and 1958 and by another 4 percent between 1958 and 1967. A comparison with the first four measures indicates that the relative increase in personal income taxes on wages and salary was responsible for the increase in these measures after 1958.

The fourth concept is the ratio of property income net of taxes to capital stock (line 6). It declined by 28 percent between 1947 and 1958 and by another 9 percent between 1958 and 1967. Its more precipitous decline compared to the preceding measure was due to the faster growth in property income tax receipts than property income over the period. The fifth concept is the ratio of after-tax corporate profits (including dividend payments) to capital stock (line 7). It fell by 47 percent between 1947 and 1958 and then increased by 28 percent between 1958 and 1967, for a net change of minus 33 percent. Its rise after 1958 reflects the relative increase in corporate profits in comparison to other forms of property income.

The sixth concept is Nordhaus' "genuine rate of return" (line 8). The numerator includes after-tax corporate profits, corporate interest payments, capital gains on net capital stock, an inventory valuation adjustment, and a correction for the capital consumption allowance, while the denominator is net capital stock. Nordhaus' estimates are uniformly higher than line 7, primarily because of his inclusion of interest payments on corporate debt and capital gains in the return. However, the trends were similar. The genuine rate of return fell by 44 percent between 1947 and 1958 and rose by 63 percent between 1958 and 1967.[21]

The last measure is Feldstein and Summers' net rate of return (line 9). The numerator includes before-tax profits and interest payments and the denominator includes net fixed capital stock, inventories, and land. Feldstein and Summers' estimated rates of return are uniformly higher than Nordhaus', because of their inclusion of before-tax, instead of after-tax, corporate profits. The trends were quite similar: a 36 percent fall between 1947 and 1958 followed by a 40 percent increase between 1958 and 1967, for a net change of minus 11 percent.

C. *Productivity and Wage Effects*

Nordhaus argues that over the postwar period sizeable changes in the relative rental rates on capital and labor occurred but induced very little substitution of capital for labor. As a result, the decline in the rental rate of capital relative to the wage rate over the period caused a decline in capital's share (particularly between 1966 and 1973).

An alternative explanation comes from the Marxian model, in which the movement of

[20]Prices of production were also computed where the general rate of profit was defined as the ratio of profits to the fixed capital stock. The trend of this general rate of profit over the period was almost identical to that of π (line 1).

[21]The net change over the period was minus 9 percent, less than line 7. The reason was the increased share of debt in corporate securities and thus the rise in interest payments relative to profits.

TABLE 3—AVERAGE LABOR CONTENT BY COMPONENT IN CONSTANT (1958) PRICES[a]

	1947	1958	1963	1967
Unweighted	.186	.140	.119	.104
Gross Domestic Output (X)	194	.146	.123	.108
Circulating Capital (aX)	.194	.140	.115	.100
Fixed Capital (kX)	.162	.141	.119	.113
Worker Consumption (M)	.199	145	.124	.107
Surplus Consumption (Y)	.190	.164	.142	.131

[a]Statistics are in man-years per thousand dollars in 1958 prices.

the rate of profit depends on relative productivity increases in the production of different components of gross output as well as the real magnitudes of these components. The labor value of a commodity is the direct plus indirect labor required to produce it. Therefore, a decline in the labor value of a commodity indicates that less labor is required to produce it or, alternatively, that "total labor productivity" has increased (see Morishima, ch. 1). Over a period of time, the increase in labor productivity may be greater for the production of some commodities than for that of others. This may result in a relatively greater absorption of labor value over time by some components of gross output than by others.

Table 3 shows the average labor values of different components of gross output and of the capital stock over the 1947–67 period. In unweighted terms, the average labor value of output declined by 44 percent between 1947 and 1967. Labor productivity thus increased at an average (compounded) rate of 2.9 percent a year over the entire period with the average rate increasing from 2.6 percent per year in the 1947–58 period to 3.3 percent a year in the 1958–67 period. Weighted by gross domestic output, the average labor value of output also declined by 44 percent over the entire period. The annual rate of productivity increase was also 2.9 percent for the entire period, 2.6 percent for the 1947–58 period and 3.3 percent for the 1958–67 period. The average labor content of gross domestic output was consistently higher than the unweighted figure, indicating that larger sectors had higher total labor requirements. The average annual rate of productivity increase in the production of circulating capital (or intermediate output) over the 1947–67 period was 3.3 percent, the highest of any of the four components. The average annual rate was 2.9 percent in the 1947–58 period and 3.7 percent in the 1958–67 period. Its labor content was lower than that of gross domestic output and that of any of the other components in all years except the first. The average annual rate of productivity increase was second highest for worker consumption at 3.1 percent for the whole period, 2.8 percent for the 1947–58 period, and 3.3 percent for the 1958–67 period. The average labor content of this component was approximately equal to that of gross domestic output in all years except the first. The average annual rate of productivity increase was third highest for surplus consumption at 1.8 percent for the entire period, 1.3 percent for the earlier period, and 2.5 percent for the later period. Its average labor content was the greatest in all years except the first. The lowest rate of productivity increase occurred in the production of fixed capital at 1.8 percent for the entire period, 1.3 percent for the 1947–58 period, and 2.4 percent for the 1958–67 period.

The effect of these differential gains in labor productivity can be seen in the movement of constant capital, variable capital, and surplus value over time. Let us decompose these movements as follows:

$$C^2 - C^1 = \lambda^2 R^2 - \lambda^1 R^1$$

where superscripts 1 and 2 refer to periods 1 and 2, respectively, and $R = (a + k)X$. Let

TABLE 4—THE PERCENTAGE DECOMPOSITION OF THE MOVEMENTS OF CONSTANT CAPITAL, SURPLUS VALUE, AND VARIABLE CAPITAL OVER TIME

	1947–58	1958–63	1963–67	1947–67
Constant Capital				
Increased Labor Productivity $((\Delta\lambda)R^1/C^1)$	−18.8	−15.8	−8.0	−46.4
Growth in Physical Capital $(\lambda^2\Delta R/C^1)$	33 7	17.3	18.8	75.6
Net Increase $((C^2 - C^1)/C^1)$	14.9	1.2	10.8	29.2
Surplus Value				
Increased Labor Productivity $((\Delta\lambda)Y^1/S^1)$	−15.0	−11.9	−7.5	−37.7
Growth in Surplus Consumption $(\lambda^2\Delta Y/S^1)$	27.5	22.1	18.2	75.0
Net increase $((S^2 - S^1)/S^1)$	12.6	10.2	10.7	37.2
Variable Capital				
Growth in Employment $(\lambda^1 m^1\Delta N/V^1)$	10.5	7.5	9.8	29.7
Increased Labor Productivity $((\Delta\lambda)m^1N^2/V^1)$	−26 7	−13.9	−13.4	−57.0
Growth in Consumption per Worker $(\lambda^2N^2\Delta m/V^1)$	24 6	11.1	12.4	50.7
Net increase $((V^2 - V^1)/V^1)$	8.4	4 7	8.7	23.4

Note: All numbers are percentages

p_1^2 be the vector of sectoral price indices from period 1 to period 2. Then,

$$C^2 - C^1 = \lambda^2 R^2 - \lambda^2(p_1^2 R^1) + (\lambda^2 p_1^2)R^1 - \lambda^1 R^1$$

$$C^2 - C^1 = (\Delta\lambda)R^1 + \lambda^2\,\Delta R$$

where $\Delta\lambda$ is the change in labor values in year 1 prices and ΔR the change in physical capital in year 2 prices. Thus,

$$C^2 = C^1 + (\Delta\lambda)R^1 + \lambda^2\,\Delta R$$

Likewise,

$$S^2 = S^1 + (\Delta\lambda)Y^1 + \lambda^2\,\Delta Y$$

With some additional manipulation, moreover, it can be shown that

$$V^2 = V^1 + \lambda^1 m^1 \Delta N + (\Delta\lambda)m^1 N^2 + \lambda^2 N^2 \Delta m$$

where ΔN is the change in employment between period 1 and period 2.

In Table 4 the movement of constant capital and surplus value over time is decomposed into a productivity effect and a growth effect and the movement of variable capital into an employment effect, a productivity effect, and a growth effect. For example, constant capital (in labor value terms) increased by 15 percent between 1947 and 1958. This was due to a 34 percent increase in the quantity of circulating capital and capital stock and a 19 percent depreciation in labor values from increased productivity. In the case of constant capital, the average annual rate of productivity increase and the average annual rate of growth of the quantity of physical capital were greater in the 1958–67 period than in the 1947–58 period. Since these are offsetting tendencies, the average (compounded) annual rate of growth of constant capital was approximately the same in the two periods, at 1.3 percent. Over the entire 1947–67 period, constant capital increased by 29 percent from a 76 percent increase in physical capital and a 46 percent increase in labor productivity.

In the case of surplus value, the average annual rate of productivity increase and of the growth in (real) surplus consumption was larger in the 1958–67 period than in the preceding one. However, in net, the average annual rate of growth of surplus value was considerably higher in the second period than in the first—2.1 percent as against 1.1 percent. Over the entire 1947–67 period, the growth in surplus consumption was 75 percent, almost the same as that of physical capital. Because the growth in labor productivity was considerably smaller in the production of surplus consumption goods than in the production of physical capital, the net increase in surplus value was considerably larger than that of constant capital.

In the case of variable capital, the average annual growth rates for employment, labor

productivity, and (real) consumption per worker were all greater in the post-1958 period than in the pre-1958 period. In net, variable capital increased at a faster rate in the later period (1.4 percent per year) than in the earlier one (0.7 percent per year). Over the entire period, the net increase in variable capital was 23 percent, the lowest of the three components. This was primarily due to the 57 percent increase in labor productivity in the production of worker consumption goods, the highest of the three components, since the 80 percent (29.7 + 50.7) increase in total worker consumption exceeded the percentage growth of the other two components. Moreover, the increase in labor productivity more than offset the growth in consumption per worker over the entire period and in each of the three subperiods. In fact, variable capital per worker declined by 5 percent over the 1947–67 period.[22] The only force leading to a positive rate of growth in variable capital in this period was thus the increase in employment.

The relative movements of the average labor content and the real magnitudes of these three components translate directly into changes in the *value* rate of profit. Moreover, with a somewhat different technique, the movement of the general rate of profit can be shown to depend on productivity changes and changes in real magnitudes. In particular, the movement of the general rate of profit can be decomposed into a technological change effect and a real wage effect (see Table 5). Along each row, it is assumed that consumption per worker (m) remains constant. Down each column, it is assumed that the technology (as defined by a, k, and l) remains constant. Between 1947 and 1958, the general rate of profit dropped by 8 percent from .139 to .128. If technology had remained constant in this period but consumption per worker had increased to its 1958 level (and composition), the general rate of profit would have dropped by 34 percent to .092. If consumption per worker had remained constant but technology had shifted to its 1958 structure, the general

TABLE 5—THE GENERAL RATE OF PROFIT WITH CONSTANT TECHNOLOGY AND CONSTANT CONSUMPTION PER WORKER

Constant Consumption Per Worker	Constant Technology (a,k,l) 1947	1958	1967
1947	.139	.161	–
1958	.092	.128	.167
1967	–	.092	.138

rate of profit would have risen by 16 percent to .161. Between 1958 and 1967, the general rate of profit increased by 8 percent from .128 to .138. If technology had remained constant, the general rate of profit would have fallen by 28 percent to .092. If consumption per worker had remained constant, the general rate of profit would have risen by 30 percent to .167.

Thus, the wage effects on the rate of profit were almost identical in the two periods, at annual (compounded) rates of −3.7 percent in the 1947–58 period and −3.6 percent in the 1958–67 period. Moreover, this is true despite the fact that real earnings per worker increased by 2.7 percent per year in the earlier period and 3.0 percent in the later period. The technological change effect, however, was considerably stronger in the later period, with an annual (compounded) rate of 3.0 percent in comparison to a rate of 1.3 percent in the 1947–58 period.

This result is consistent with the findings above on the relative magnitudes of productivity movements over this period (see Tables 3 and 4). The rate of growth of labor productivity for all components of gross output was greater in the 1958–67 period than in the 1947–58 period. The rate of growth in real earnings per worker was also greater in the later period than in the earlier one. However, in the 1947–58 period, the rate of labor productivity increase was less than the rate of growth of the real wage (2.55 percent per year compared to 2.72), causing the general rate of profit to fall by 8 percent. In the 1958–67 period, the rate of labor productivity increase was greater than the rate of growth of the real wage (3.29 percent per year

[22]This is equivalent to a rise in the rate of surplus value (see equation (10)).

compared to 3.02), causing the general rate of profit to rise by 8 percent. Thus, the movement of the general rate of profit during this period depended primarily on the relative increase in the real wage and labor productivity.[23]

III. Conclusion

Though Marx's law of the tendency of the rate of profit to fall is theoretically unsound and though there is no empirical support for it during my period of investigation, Marx's formulation of the problem is still useful in isolating the key determinants of the movement of the rate of profit. In particular, my analysis indicates that its movement was primarily a result of the counteracting effects of labor productivity changes and real wage changes. In fact, not only the general rate of profit but all the measures I used fell between 1947 and 1958, a period when real wage increases outweighed increases in labor productivity. In addition, all but two measures rose between 1958 and 1967, a period when the converse was true. (The two that did not rise were measures in which some portion of tax receipts were excluded in the definition of the gross profit.) Moreover, a surprising finding, similar to that of Nordhaus, is that changes in the capital-labor ratio (the technical composition of capital) had little to do with movements in the rate of profit. The reason is that increases in the capital-labor ratio were almost entirely offset by corresponding increases in labor productivity. Future work might be done on the 1966–74 period in the United States, when the rate of profit fell sharply, and the 1974–76 period, when it rose, to verify these conclusions when the necessary data become available.

[23]Some caution should be exercised in interpreting this result, since the analysis depends to a large extent on the simplifying assumptions that were made in constructing the modified Leontief framework. In particular, our definition of the gross surplus as total value-added less the necessary costs of reproducing the labor force may have resulted in an overstatement of the importance of changes in the real wage on the rate of profit. Moreover, by assuming homogeneous labor, an annual turnover rate in all sectors, and no unproductive activity, we may have overstated the importance of the productivity effect.

REFERENCES

W. Baumol, "The Transformation of Values: What Marx 'Really' Meant," *J. Econ. Lit.*, Mar. 1974, *12*, 51–62.

J. Christiansen, "Marx and the Falling Rate of Profit," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 20–6.

S. Dresch and R. Goldberg, "IDIOM," Nat. Bur. Econ. Res. work. paper, New York, Feb. 1975.

M. Feldstein and L. Summers, "Is the Rate of Profit Falling?," *Brookings Papers*, Washington 1977, *1*, 211-27.

John Kendrick, *The Formation and Stocks of Total Capital*, New York 1976.

O. Kýn, B. Sekerka, and L. Hejl, "A Model for the Planning of Prices," in Charles H. Feinstein, ed., *Socialism, Capitalism, and Economic Growth*, Cambridge 1967.

J. Lecaillon and D. Germidis, "Economic Development and the Wage Share in National Income," *Int. Labor Rev.*, May 1975, *3*, 393–409.

Karl Marx, *Capital*, New York 1967.

Michio Morishima, *Marx's Economics*, Cambridge 1973.

——— **and F. Seton,** "Aggregation in Leontief Matrices and the Labour Theory of Value," *Econometrica*, Apr. 1961, *29*, 203–20.

W. Nordhaus, "The Falling Share of Profits," *Brookings Papers*, Washington 1974, *1*, 169–208.

N. Okishio, "Measurement of the Rate of Surplus Value," *Hitotsubashi J. Econ.*, Oct. 1959, *10*, 297–303.

J. Roemer, "Some Clarifications of the 'Tendency of the Rate of Profit to Fall'," mimeo, work. paper no. 64, Univ. of California-Davis, June 1976.

F. Seton, "The Transformation Problem," *Rev. Econ. Stud.*, June 1957, *24*, 149–60.

P. Samuelson, "Understanding the Marxian Notion of Exploitation," *J. Econ. Lit.*, June 1971, *9*, 399–431.

Akira Takayama, *Mathematical Economics*, Hinsdale 1974.

E. **Wolff**, "The Rate of Surplus Value in Puerto Rico," *J. Polit. Econ.*, Oct. 1975, *83*, 935–49.

———, (1977a) "Capitalist Development, Surplus Value, and Reproduction: An Empirical Examination of Puerto Rico," in Jesse Schwartz, ed., *The Subtle Anatomy of Capitalism*, Santa Monica 1977.

———, (1977b) "Unproductive Labor and the Rate of Surplus Value in the United States," in Paul Zarembka, ed., *Research in Political Economy*, Greenwich 1977.

U.S. Bureau of Labor Statistics, *The Structure of the U.S. Economy in 1980 and 1985*, Washington 1975.

U.S. Council of Economic Advisers, *Economic Report of the President*, Washington 1972.

Engineering and Econometric Interpretations of Energy-Capital Complementarity

By ERNST R. BERNDT AND DAVID O. WOOD*

The estimation of Hicks-Allen substitution elasticities among energy and nonenergy inputs in manufacturing processes has been the subject of several recent econometric studies. Perhaps the most interesting result of these studies has been the apparent contradictory evidence regarding substitution possibilities between energy and capital. The principal results may be summarized as follows:

(i) Using time-series data on capital (K), labor (L), energy (E), and other nonenergy intermediate goods (M) for *U.S.* manufacturing, the authors (1975), and Berndt and Dale W. Jorgenson find E–K complementarity. Similar results are obtained by Melvyn Fuss using *KLEM* time-series data for Canadian manufacturing pooled by region, by Jan R. Magnus using *KLE* time-series data for Dutch manufacturing, and by Paul Swaim and Gerhard Friede with *KLEM* time-series data for the West German industrial sector.

(ii) In contrast, based on *KLE* time-series data for manufacturing pooled by Organization for Economic Cooperation and Development (*OECD*) countries, James M. Griffin and Paul R. Gregory as well as Robert S. Pindyck report E–K substitutability. Similarly, using cross-section data on K, two types of L, and three types of E by state for eight 2-digit SIC manufacturing industries, Robert Halvorson and Jay Ford find either significant E–K substitutability or insignificant complementarity. Makota Ohta has employed the hedonic approach with data on *U.S.* electric utility purchases and energy efficiency characteristics of boilers and turbogenerators, and finds a significant tradeoff between initial capital costs and improvements in energy efficiency. Thomas Cowing reports similar results.

The econometric studies finding evidence of E–K substitutability are apparently consistent with and supported by engineering analyses of E conservation potential due to E–K substitution. A number of such studies have shown that the average energy efficiency of existing plant and equipment is a small fraction (5–10 percent) of the maximum possible efficiency based on the Second Law of Thermodynamics.[1] More detailed engineering studies have estimated increases in energy efficiency available through increased investment either by retrofitting existing plant and equipment or through new engineering designs.[2] As one example, Gyftopoulos and Widmer calculate corresponding percentage changes in fuel efficiency and the initial capital cost for selected industrial processes, and find significant tradeoffs.

Conflicting evidence from econometric studies regarding E–K substitution might be explained by a number of reasons including differing data sets and approaches to measuring input quantities and prices, varied treat-

*Associate professor of economics, University of British Columbia , on sabbatical leave at M.I.T., and senior lecturer, M.I.T. Sloan School of Management, respectively. This paper is a considerably shortened version of our 1977 paper. Our interest in this topic was stimulated by discussions with colleagues on the National Academy of Sciences – National Research Council Committee on Nuclear and Alternative Energy Systems (*CONAES*). We are indebted to Cathy White, M. S. Khaled, and Supriya Lahiri for exceptional research assistance, and to numerous colleagues for helpful comments.

[1]See, for example, Elias P. Gyftopoulos, Lazaros Lazaridis, and Thomas Widmer and the references cited therein. For a discussion on the relationship between economic and thermodynamic measures of energy efficiency, see Berndt (1978)

[2]It is not possible for us to provide an exhaustive set of literature references. In addition to references cited in the much publicized studies of the Ford Foundation Energy Policy Project and the U.S. Federal Energy Administration *Project Independence* report, see U.S. Office of Emergency Preparedness, U.S. National Academy of Sciences, National Petroleum Council (pp. 24–31), Eric Hirst, Hirst et al., Hirst and Janet Carney, Lee Schipper and Joel Darmstadter, Darmstadter et al., and Robert Socolow.

ments of excluded inputs and distinctions between short- and long-term elasticities. Although for production models involving more than two inputs there is no a priori reason to hypothesize either E–K substitutability or complementarity, E–K substitutability appears to be supported by the weight of the engineering evidence. This evidence might cast serious doubt on the reasonableness of the E–K complementarity econometric results unless it could be demonstrated that all three sets of results can be reconciled within a common analytical framework.

Our purpose in this paper is to provide such a reconciliation and interpretation of engineering and econometric studies. In Section I we develop an analytical framework which reconciles the engineering evidence with the possibility of E–K complementarity. In Sections II and III we present empirical evidence reconciling the seemingly disparate econometric results of E–K complementarity and substitutability.

I. Theoretical Foundations

We begin by considering a positive, twice-differentiable strictly quasi-concave production function with n inputs:

$$(1)\quad Y = F(x) = F(x_1, x_2, x_3, \ldots, x_n),\ x_i > 0$$

relating maximum possible output Y obtainable from any given set of inputs. The set of n inputs is denoted $N = [1, \ldots, n]$, and $F_i \equiv \partial F/\partial x_i$, $F_{ij} \equiv \partial^2 F/\partial x_i \partial x_j$. We partition the set N into r mutually exclusive and exhaustive subsets $N_1, N_2, \ldots, N_r$, a partition we call R. The production function (1) is homothetically weakly separable with respect to the partition R if the marginal rate of substitution between any two inputs x_i and x_j from any subset N_m, $m = 1, \ldots, r$ is independent of output and the quantities of inputs external to N_m, i.e.,

$$(2)\quad \frac{\partial (F_i / F_j)}{\partial x_k} = 0 \quad i, j \in N_m \text{ and } k \notin N_m$$

Homothetic weak separability with respect to the partition R is necessary and sufficient for the production function (1) to be of the form $F^*(X^1, X^2, \ldots, X^r)$ where X^m is a positive, strictly quasi-concave homothetic production subfunction of only the elements within N_m, i.e.,

$$(3)\quad X^m = f_m(x_i),\ i \in N_m,\ m = 1, \ldots, r$$

Engineering process analysis studies which focus only on a subset of the inputs and ignore all other inputs are appropriate if the subset of inputs is homothetically weakly separable from all others. When f_m is linear homogeneous in x_i, X^m is called a consistent aggregate index of the inputs in N_m. Hence econometric studies utilizing input aggregates such as labor (or energy) are valid if the components of labor (energy) are homothetically weakly separable from all other nonlabor (nonenergy) inputs.[3]

Both economists and engineers analyze production processes with the objective of minimizing production costs. When Y and input prices are exogenous, the cost-minimization assumption implies the existence of a cost function dual to $F(x)$,

$$(4)\quad C = G(Y, p_1, p_2, \ldots, p_n)$$

relating the minimum possible total cost of producing Y to the positive input prices, technological parameters, and Y. Moreover, when the production function is homothetically weakly separable in the partition R, the corresponding dual cost function has the same partition in input prices, that is, the dual cost function is

$$(5)\quad C = G^*(Y, P^1, P^2, \ldots, P^m)$$

where the input price aggregates P^m are positive, strictly quasi-concave homothetic functions of only the elements in N_m, i.e.,

$$(6)\quad P^m = g_m(p_i),\quad i \in N_m,\ m = 1, \ldots, r$$

Hereafter we assume that the f_m and g_m functions are linear homogeneous and that economic profits are zero. Thus the "output

[3]Further discussion of separability is found in Berndt and Laurits R. Christensen.

price" P^m of the cost subfunctions in (6) is equal to the unit cost of producing X^m.

As an example, let the production function be

$$(7) \qquad Y = F(K, L, E, M)$$

where K, L, E, and M are input aggregates of capital services, labor, energy, and nonenergy intermediate materials. We now specify a two-input linearly homogeneous weakly separable subfunction with only two inputs—K and E—which produces the "utilized capital" output K^*,

$$(8) \qquad K^* = f(K, E)$$

For example, consider the production of industrial process steam of given specified physical characteristics. In such a context utilized capital services K^* refers to the quantity of steam produced per unit of time using capital (in the form of boilers and supporting combustion equipment) and fuel inputs. This assumption of a separable utilized capital subfunction implies that the optimal E/K ratios within (8) depend solely on P_K, P_E and not on the other input prices P_L, P_M or the level of gross output Y. Dual to the master production function (7) is the master cost function

$$(9) \qquad C = G^*(Y, P_K, P_L, P_E, P_M)$$

The corresponding dual utilized capital cost subfunction is separable in the same input partition,

$$(10) \qquad C_{K^*} = G(K^*, P_K, P_E)$$

where $C_{K^*} = P_K K + P_E E$.

We now turn to a discussion of price elasticities. Suppose a researcher seeks to measure cross-price elasticities between two inputs x_i and x_j within the same input subset N_m. We define the *gross price elasticity* ϵ^*_{ij} as the logarithmic derivative $\partial ln x_i/\partial ln P_j (i, j \in N_m)$, where X^m, the output of the weakly separable subfunction (3) is held constant, all $x_k (k \in N_m)$ adjust to their optimal levels, but all $x_q (q \in N_m)$ are held fixed. We define the *net price elasticity* ϵ_{ij} as the derivative $d\, ln\, x_i/d\, ln\, p_j (i, j \in N_m)$ where instead of X^m, the "master" output Y is held fixed at $Y = Y_0$ and all other inputs—not just those in N_m—adjust to their new cost-minimizing levels. The gross price elasticity ϵ^*_{ij} is conditional on fixed X^m (say, $X^m = X^m_0$), whereas the net price elasticity ϵ_{ij} permits X^m to respond to the change in p_j. The relationship between ϵ^*_{ij} and $\epsilon_{ij} (i, j \in N_m)$ is

$$(11) \qquad \left.\frac{d\, ln\, x_i}{d\, ln\, p_j}\right|_{Y=Y_0} = \left.\frac{\partial\, ln\, x_i}{\partial\, ln\, p_j}\right|_{X^m=X^m_0} + \left.\frac{\partial\, ln\, x_i}{\partial\, ln\, X^m} \cdot \frac{\partial\, ln\, X^m}{\partial\, ln\, P^m} \cdot \frac{\partial\, ln\, P^m}{\partial\, ln\, p_j}\right|_{Y=Y_0}$$

which, under the assumption of linear homogeneity in (3) reduces to

$$(12) \qquad \epsilon_{ij} = \epsilon^*_{ij} + S_{jm}\epsilon_{mm}, \quad i, j \in N_m$$

where S_{jm} is the cost share of the jth input in the total cost of producing X^m and ϵ_{mm} is the own-price elasticity of demand for X^m along a $Y = Y_0$ isoquant. Thus the net effect on the derived demand for x_i, given a change in the price of $x_j (i, j \in N_m)$ is the sum of a gross effect conditional on the fixed input aggregate X^m plus the cost share S_{jm} times the own-price elasticity ϵ_{mm}. Since ϵ_{mm} is negative and S_{jm} is positive, it follows that $S_{jm}\epsilon_{mm} < 0$ and that $\epsilon_{ij} < \epsilon^*_{ij}$. Hereafter we define $S_{jm}\epsilon_{mm}$ as the expansion elasticity.[4] Notice that when the negative expansion elasticity dominates the positive gross price elasticity, i and j are gross substitutes ($\epsilon^*_{ij} > 0$), but net complements ($\epsilon_{ij} < 0$).

In the context of our utilized capital subfunction specification net and gross price elasticities between E and K are

$$(13) \qquad \epsilon_{ij} = \epsilon^*_{ij} + S_j\epsilon_{mm}, \quad i, j = K, E, \quad m = K^*$$

where S_K and S_E are the cost shares of K and E in the total cost of producing the utilized capital output K^* and $\epsilon_{K^*K^*}$ is the price elasticity of demand for utilized capital services K^*. Equation (13) indicates, for example, that the net price elasticity ϵ_{KE} along a $Y = Y_0$ isoquant equals the positive gross substitution

[4]The term "expansion elasticity" is due to Leif Johansen (pp. 124–26); it characterizes movement along an expansion path. The term "scale elasticity" used in our 1977 paper has been discarded in order to avoid possible confusion regarding returns to scale.

elasticity ϵ^*_{KE} along a $K^* = K^*_0$ isoquant plus the cost share of energy in the K^* subfunction times the price elasticity of demand for K^*. We interpret much of the recent engineering literature dealing with possibilities for energy conservation as focusing on real world possibilities for movement along the utilized capital isoquant; in brief, this engineering-economic literature focuses on the gross price elasticity ϵ^*_{KE}, not the net price elasticity ϵ_{KE}. The analyses in the engineering-economic conservation literature illustrate that a given amount of utilized capital services can be produced with less E but more K. Since this literature typically deals with only two inputs, E–K complementarity is not possible.

We now provide a geometric interpretation of the E–K relationship. First, we specify another weakly separable linearly homogeneous production subfunction with two inputs,

$$L^* = h(L, M) \tag{14}$$

where L^* is the output of the labor-materials production subfunction. The master production function can then be written as $Y = F(K, L, E, M) = F^*(K^*, L^*)$, and the dual master cost function as $C = G(Y, P_K, P_L, P_E, P_M) = G^*(Y, P_{K^*}, P_{L^*})$ where P_{K^*} and P_{L^*} are unit costs from the cost subfunctions:

$$\begin{aligned} P_{K^*} &= C_{K^*}/K^* = g^*(P_K, P_E), \\ P_{L^*} &= C_{L^*}/L^* = h^*(P_L, P_M) \end{aligned} \tag{15}$$

and where $C_{L^*} = P_L L + P_M M$.

Consider the following example. Suppose that a cost-minimizing competitive firm was in equilibrium producing gross output $Y = Y^*_1$. Given the original input prices P_{K^*} and P_{L^*}, the slope of the iso-cost line AA' in Figure 1 is $-P_{L^*}/P_{K^*}$, and the firm minimizes costs of producing Y^*_1 at O_1 using K^*_1 units of utilized capital and L^*_1 units of the labor-materials composite. Given the original prices P_K and P_E as reflected in the iso-cost line BB' in Figure 2, the firm produces the K^*_1 output at O_2 using K_1 units of capital and E_1 units of energy; similarly, given P_L and P_M as reflected in the iso-cost line CC' in Figure 3, the L^*_1

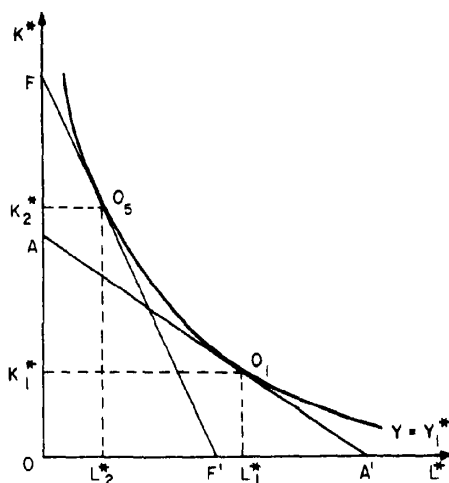



FIGURE 1. MASTER FUNCTION ISOQUANT

output is produced at O_3 using L_1 units of labor and M_1 units of materials.

Next, assume the federal government introduces investment incentives that lower P_K. The total effect on the elasticity of demand of capital ϵ_{KK}, and on the demand for energy ϵ_{EK}, consists of two components as shown in (13). First, holding fixed the output of the utilized capital subfunction $K^* = K^*_1$ the steeper iso-cost line DD' in Figure 2 (due to the lower P_K) indicates that demand for capital would increase from K_1 to K_2, and that demand for energy would fall from E_1 to E_2; these gross substitution effects are represented in (13) by ϵ^*_{KK} and ϵ^*_{EK}, respectively.

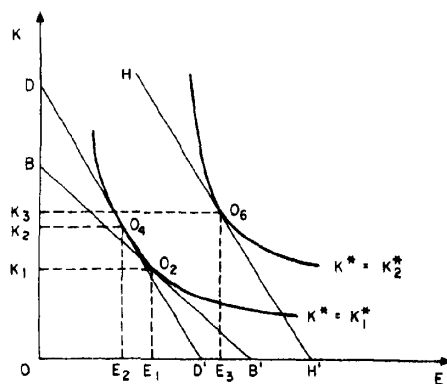



FIGURE 2. K^* ISOQUANT

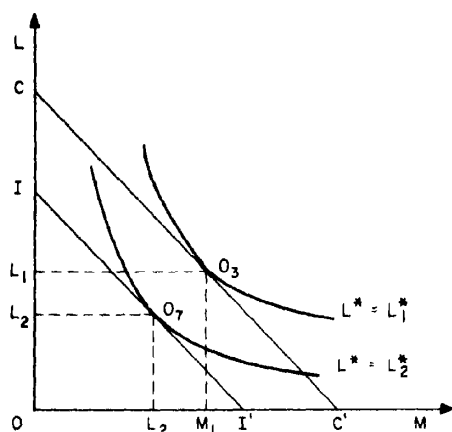



FIGURE 3. L^* ISOQUANT

Second, since the investment incentives decrease P_K, this reduces the cost C_K^* of producing utilized capital services, and by (15) lowers P_K^*. This changes the iso-cost line in Figure 1 from AA' to a steeper iso-cost line FF', and results in a new cost-minimizing equilibrium at O_5 where derived demand for utilized capital increases from K_1^* and K_2^*, while demand for L^* falls from L_1^* to L_2^*. This results in an outward shift of the K^* isoquant, as shown in Figure 2, increasing derived demand both for capital and energy. At the new equilibrium O_6 this expansion effect increases derived demand for capital and energy from K_2 to K_3 and from E_2 to E_3. For capital, the gross substitution effect (K_1 to K_2) and the expansion effect (K_2 to K_3) reinforce each other, but for energy the two effects work in opposite directions; the gross substitution effect decreases energy demand from E_1 to E_2, whereas the expansion effect increases demand from E_2 to E_3. Note that E_3 is larger than E_1. In the particular example of Figure 2, the expansion effect $S_K\epsilon_{K^*K^*}$ dominates the gross substitution effect ϵ_{EK}^*. Thus, in this example E and K are gross substitutes ($\epsilon_{EK}^* > 0$) but net complements ($\epsilon_{EK} < 0$). Note that in this context net E–K complementarity implies that the investment incentives contribute to increased (not reduced) energy demand. Whether net K–E substitutability or complementarity exists depends then on whether the gross substitution effect or the expansion effect is dominant. That, of course, is an empirical issue.

II. Empirical Analysis

In order to implement the separable model empirically, we must specify functional forms for the master function and for the K^* and L^* subfunctions. For convenience, we utilize the dual cost functions and employ the flexible translog specification

$$(16)\quad \ln P^m = \ln(C_m/m) = \ln \delta_m + \sum_i \alpha_i \ln P_i + \frac{1}{2}\sum_i \sum_j \gamma_{ij} \ln P_i \ln P_j,$$
$$\gamma_{ij} = \gamma_{ji}, \quad i,j,m = K, E, K^*; L, M, L^*$$

where

$$(17)\quad \sum_i \alpha_i = 1.0, \quad \sum_j \gamma_{ij} = 0, \quad i,j = K,E; L,M$$

To complete our empirical model specification, we assume the master production function is the familiar strongly separable linearly homogeneous Cobb-Douglas function with Hicks-neutral constant exponential technical change. The dual master unit cost function is then written as

$$(18)\quad \ln(C/Y) = \ln \beta_o + \alpha_t t + \beta_{K^*} \ln P_{K^*} + \beta_{L^*} \ln P_{L^*}$$

where P_{K^*} and P_{L^*} are defined as equalling the unit cost of K^* and L^* (see (15)), t represents time, and $\beta_{K^*} + \beta_{L^*} = 1$.

Substituting the translog subfunctions (16) into (18), we write the master cost function in terms of the separable subfunction prices and parameters:

$$(19)\quad \ln(C/Y) = \ln \alpha_o + \alpha_t t + \sum_i \beta_i \ln P_i + \frac{1}{2}\sum_i \sum_j \beta_{ij} \ln P_i \ln P_j,$$

where $\beta_{ij} = \beta_{ji}, \quad \sum_i \beta_{ij} = \sum_j \beta_{ij} = 0,$
$\sum_i \beta_i = 1, \quad i,j = K,L,E,M$

and

$$(20) \quad \ln \alpha_o = \ln \beta_o + \beta_{K^*} \ln \delta_{K^*} + \beta_{L^*} \ln \delta_{L^*}$$

where

$$\beta_{KL} = \beta_{KM} = \beta_{LE} = \beta_{EM} = 0$$
$$\beta_{KK} = \beta_{K^*}\gamma_{KK} \qquad \beta_{KE} = \beta_{K^*}\gamma_{KE}$$
$$\beta_{LL} = \beta_{L^*}\gamma_{LL} \qquad \beta_{LM} = \beta_{L^*}\gamma_{LM}$$
$$\beta_K = \beta_{K^*}\alpha_K \qquad \beta_L = \beta_{L^*}\alpha_L$$
$$\beta_M = \beta_{L^*}\alpha_M \qquad \beta_{EE} = \beta_{K^*}\gamma_{EE}$$
$$\beta_{MM} = \beta_{L^*}\gamma_{MM} \qquad \beta_E = \beta_{K^*}\alpha_E$$

Based on the master cost function specification (18)–(20) we compute Hicks-Allen partial elasticities of substitution σ_{ij} and price elasticities ϵ_{ij} along a gross output isoquant as

$$(21) \quad \sigma_{ij} = \frac{\beta_{ij} + S_i S_j}{S_i S_j} \qquad i, j = K, L, E, M, i \neq j$$

$$\sigma_{ii} = \frac{\beta_{ii} + S_i^2 - S_i}{S_i^2} \qquad i = K, L, E, M$$

and

$$(22) \quad \epsilon_{ij} = S_j \sigma_{ij} \qquad i, j = K, L, E, M$$

where the S_i are the cost shares of the ith input in the total cost C of producing gross output, obtained by logarithmically differentiating (19) and using Shephard's Lemma:

$$(23) \quad S_i = \frac{\partial \ln C}{\partial \ln P_i} = \frac{P_i X_i}{C} = \beta_i + \sum_j \beta_{ij} \ln P_j, \qquad i, j = K, L, E, M$$

The price elasticities in (22) are net price elasticities. We can of course rewrite the net price elasticity ϵ_{ij} in terms of the gross price elasticity ϵ_{ij}^* and the expansion elasticity. First we write the cost shares within the K^* and L^* subfunctions as

$$(24) \quad S_{im} = \frac{\partial \ln (C_m/m)}{\partial \ln P_i} = \frac{P_i X_i}{C_m} = \alpha_i + \sum_j \gamma_{ij} \ln P_i, \quad i, j, m = K, E, K^*; L, M, L^*$$

and note that

$$(25) \quad S_{im} = \beta_m S_i, \quad i = K, E, \quad m = K^*; \quad i = L, M, m = L^*$$

Then, in the context of E–K net price elasticities, we obtain

$$(26) \quad \epsilon_{ij} = \epsilon_{ij}^* + S_{jm}\epsilon_{mm} \qquad i, j, m = K, E, K^*; \quad L, M, L^*$$

An interesting feature of (20) is that these constraints constitute a set of testable parametric restrictions on the more general four input *KLEM* translog unit cost function (19) with Hicks-neutral constant exponential technical change. In our 1975 paper we have called the set of restrictions (20) on (19) linear separability restrictions for $[(K, E), (L, M)]$ separability. These restrictions were not rejected with our data—annual *U.S.* manufacturing, 1947–71.[5] Furthermore, in that paper, we tested for different types of separability among *KLEM* inputs; all forms except that represented in (20) were rejected. These econometric findings therefore provide some empirical support for our simple nested model specification.[6]

We append an additive disturbance term to each of the above equations (23). Assuming that the truncated disturbance vector (with S_M deleted) is independently and identically multivariate normally distributed with mean-vector zero and constant nonsingular covariance matrix, we have estimated the S_K, S_L,

[5] In our 1975 paper we presented estimated elasticities based on iterative three-stage least squares (*I3SLS*) estimation. In this paper, we assume input prices and gross output quantity are exogenous, and estimate parameters of the share equations (23) using maximum likelihood (*ML*) procedures; the *I3SLS* and *ML* estimates are virtually identical. The likelihood ratio test statistic of the four independent restrictions in (20) using *ML* estimation is 10.326, while the .01 *chi*-square critical value is 13.277; under *I3SLS* estimation, the Wald test statistic is 9.038 and the .01 *chi*-square critical value remains 13.277.

[6] We have also tested for the validity of a related utilized capital separability specification using the three-input *KLE* data of Griffin-Gregory, which they kindly provided us. Based on their data, the likelihood ratio test statistic for the two restrictions is 3.2502, while the .01 *chi*-square critical value is 9.210. Hence, using the Griffin-Gregory *KLE* data, we cannot reject the null hypothesis of non-linear $[(K, E), L]$ separability.

and S_E translog cost-share equations (23) with the separability restrictions (20) imposed using the method of maximum likelihood.[7] Based on these ML parameter estimates, we have computed ML estimates of selected net, expansion and gross elasticities for total $U.S.$ manufacturing in the last year of our sample (1971).[8] These elasticity estimates are presented in the top panel of Table 1. There it is seen that although E and K are gross substitutes in $U.S.$ manufacturing, they are also net complements. The gross substitution effect (.133) is dominated by the expansion effect (−.462), resulting in a value of the net elasticity ϵ_{EK} of −.329. Moreover, these elasticities are significantly different from zero.

The above results were based on annual $U.S.$ manufacturing time-series data. To investigate the robustness of our net complementarity findings within the utilized capital framework, we have also estimated a slightly generalized model, due to Fuss, using a nonhomothetic translog $KLEM$ cost function based on pooled cross-section time-series data for Canadian manufacturing by region, 1961–71. Because Fuss specified a nonhomothetic translog cost function and used an error components estimation procedure (the "covariance" method), the conditions (20) for separability are not directly applicable or testable.[9]

Table 1—Net, Scale, and Gross Substitution Elasticities in Utilized Capital Model $[(K, E), (L, M)]$ Separability Restrictions Imposed U.S and Canadian Manufacturing, 1971

Net Elasticity	Gross Substitution Elasticity	Scale Elasticity	Value of Net Elasticity
$U.S.$ Manufacturing, 1971			
ϵ_{KK}	−.126	−.462	−.588
	(.024)	(.003)	(.026)
ϵ_{EE}	−.133	−.440	−.573
	(.026)	(.003)	(.024)
ϵ_{KE}	.126	−.440	−.314
	(.024)	(.003)	(.027)
ϵ_{EK}	.133	−.462	−.329
	(.026)	(.003)	(.026)
Canadian Manufacturing Ontario, 1971			
ϵ_{KK}	−.039	−.765	−.744
	(.009)	(.238)	(.238)
ϵ_{EE}	−.505	−.054	−.559
	(.115)	(.018)	(.117)
ϵ_{KE}	.039	−.054	−.015
	(.009)	(.018)	(.020)
ϵ_{EK}	.501	−.705	−.200
	(.115)	(.238)	(.264)
Canadian Manufacturing British Columbia, 1971			
ϵ_{KK}	−.121	−.664	−.705
	(.011)	(.206)	(.206)
ϵ_{EE}	−.650	−.123	−.773
	(.052)	(.038)	(.066)
ϵ_{KE}	.121	−.123	−.002
	(.011)	(.038)	(.040)
ϵ_{EK}	.650	−.664	−.014
	(.052)	(.206)	(.213)

[a]Estimated Asymptotic Standard Errors in Parentheses.

To preserve the distinguishing features of Fuss' paper—a nonhomothetic translog specification and the covariance estimation method—we proceed with separate estimation of the gross, net, and expansion elasticities as follows. First, the K^* and L^* subfunctions are again specified as in (16). Using Fuss' Canadian $KLEM$ manufacturing data, 1961–71 by region, we have estimated the

[7]The data used are those listed on our 1975 paper. The ML parameter estimates (asymptotic t-ratios) are β_{K^*} = .0983 (89.78), β_{L^*} = .9017 (823.46), α_K = .5702 (231.46), α_E = .4298 (174.48), α_L = .2800 (118.16), α_M = .7200 (303.84), $\gamma_{KK} = -\gamma_{KE} = \gamma_{EE}$ = .1851 (15.16) and $\gamma_{LL} = -\gamma_{LM} = \gamma_{MM}$ = .0868 (12.87). The implied ML estimates are β_K = .0561 (75.79), β_L = .2525 (124.42), β_E = .0423 (93.13), β_M = .6492 (352.84), $\beta_{KK} = -\beta_{KE} = \beta_{EE}$ = .0182 (15.38) and $\beta_{LL} = -\beta_{LM} = \beta_{MM}$ = .0782 (12.88). The generalized R^2 is .9844. For further details, see our 1977 paper.

[8]Since all our elasticity estimates are stable over the 1947–71 time period the year 1971 can be interpreted as representative. All our fitted shares were positive and strict quasi-concavity curvature conditions were satisfied for all years in our sample.

[9]We note that in the context of a linear homogeneous translog gross output function, specifying the K^* subfunction (10) to be translog necessarily implies that the L^* subfunction is translog and that the master function is Cobb-Douglas. However, in the more general nonhomothetic context, it is not possible to test for separability; see Charles Blackorby, Daniel Primont and Robert Russell; also Michael Denny and Fuss.

share equations (24) using the covariance method.[10] We then insert the resulting parameter estimates into (16) and form fitted data series for P_{K*} and P_{L*}. Next, following Fuss we specify the master function to be a nonhomothetic translog cost function and then estimate the resulting "master" cost-share equations using the covariance method. Finally, we compute the associated gross, expansion, and net price elasticities. In the bottom two panels of Table 1 we present 1971 estimates for two provinces, Ontario and British Columbia. For Canadian manufacturing, E and K are gross substitutes but net complements. The net substitution effect for Ontario in 1971 is negative (−.200), while for British Columbia the gross substitution and expansion effects almost offset each other, resulting in only very slight net complementarity (−.014). Unlike the *U.S.* manufacturing case, however, for Canadian manufacturing the negative net elasticity estimates are insignificantly different from zero.

In the preceding paragraphs we have shown, both analytically and empircally, that the engineering notion of E–K substitutability does not necessarily imply net E–K substitutability in the sense of Hicks-Allen. Thus the following Griffin-Gregory (hereafter, G-G) intuitive argument for E–K substitutability can be misleading: "... one might expect K and E to be *substitutes* since new equipment could be designed to achieve higher thermal efficiencies but at greater capital costs" (p. 846). G-G have, however, published econometric findings which appear to support Hicks-Allen E–K substitutability. Thus it remains to reconcile our econometric findings with those of G-G—a task to which we now turn.

Griffin-Gregory estimated a three-input (K, L, and E—but not M) translog cost function based on data for the manufacturing sector of nine industrialized *OECD* countries in four years—1955, 1960, 1965, and 1969. Although parameter estimates are *ML*, G-G do not use *ML* methods in estimating elasticities; in particular, actual rather than *ML* predicted cost shares are used in (21)–(22) to compute estimated elasticities. The difference between their procedure and the *ML* approach would be negligible if the estimated model fitted the data closely, but unfortunately this is not the case with the G-G model—especially for the United States. The generalized R^2 for the G-G preferred Model I is .41, and the difference between fitted and actual cost shares for the United States is considerable. For example, in 1965—the year for which G-G report elasticity estimates—the actual cost share of capital is .1436 while the G-G predicted share is .2205. Elasticity estimates are of course affected by this computational nuance. Griffin-Gregory (*ML*) estimates for ϵ^*_{KK} in the United States in 1965 are −.18 (−.34), for ϵ^*_{KL}, .05 (.22), and for ϵ^*_{EK}, .15 (.23). These differences in the G-G and *ML* estimates reflect the rather poor fit of the G-G model to their *U.S.* data and suggest caution in comparing their *U.S.* results with those of our 1975 paper. Furthermore, since the standard error estimates of G-G on β_{KE} are large, so are the standard errors of ϵ^*_{EK}. For example, for the United States the two standard error confidence intervals are $-.05 \leq \epsilon^*_{EK} \leq .51$ (*ML*) and $-.13 \leq \epsilon^*_{EK} \leq .43$ (G-G). We conclude that the G-G estimates are consistent with E–K complementarity.

An even stronger argument can be made that the G-G net elasticity estimates are upward biased, since they are based on a *KLE* rather than a *KLEM* model. Griffin-Gregory noted that comparison of their elasticity estimates with those of our 1975 paper might be questioned since, unlike us, for data availability reasons they omit M and assume $[(K, L, E), M]$ weak separability. This leads them to state that "... this omission may bias our findings if our weak separability assumption ... is invalid" (p. 852). However, even if the G-G weak separability assumption were valid, all their elasticity estimates would reflect gross substitution elasticities and therefore all would be upward biased estimates of net elasticities.

The reason for the upward bias is based on the argument developed in Section I. The G-G separability assumption implies that

[10]These data were kindly provided us by Fuss. Further details of the estimation procedure are presented in our 1977 paper.

$Y = F(K, L, E, M) = F^*[(K, L, E), M] = F^{**}(V, M)$, where V is the output of the $V = v(K, L, E)$ production subfunction. Clearly, the G-G estimates are gross elasticity estimates holding V fixed. Following (11), the KLE three input gross price elasticities ϵ_{ij}^* can be related to the $KLEM$ four-input net price elasticities ϵ_{ij} as

$$(27) \quad \epsilon_{EK} = \epsilon_{EK}^* + S_{KV}\epsilon_{VV}, \qquad \epsilon_{KE} = \epsilon_{KE}^* + S_{EV}\epsilon_{VV}$$

where S_{KV} and S_{EV} are the cost shares of K and E in the total cost of producing the $V = v(K, L, E)$ output and ϵ_{VV} is the price elasticity of demand for output V along a four-input $Y = F(K, L, E, M) = F^{**}(V, M)$ isoquant. Since S_{KV} and S_{EV} are positive cost shares and ϵ_{VV} is nonpositive, $\epsilon_{EK}^* \geq \epsilon_{EK}$ and $\epsilon_{KE}^* \geq \epsilon_{KE}$, implying that unless the output price elasticity $\epsilon_{VV} = 0$, the G-G estimates of net K–E price elasticities are upward biased.

The quantitative magnitude of this upward bias can be approximated as follows. Griffin-Gregory's 1965 values of S_{KV} and S_{EV} in the United States are .14 and .13. If we insert these into (13) and let the unknown ϵ_{VV} take on three alternative values, -0.5, -1.0, and -1.5, we obtain three alternative net elasticity estimates: .08, .01, and $-.06$ for ϵ_{EK} and .065, .0 and $-.065$ for ϵ_{KE}. Since these point estimates include negative values, it follows that the gross price elasticity estimates of G-G are in fact consistent with net K–E complementarity. This analytical argument applies equally to all three-input KLE studies.

We conclude that the apparently inconsistent E–K complementarity of the authors and G-G E–K substitutability econometric findings may simply be due to the fact that different elasticities are being compared; when the distinction between net and gross elasticities is acknowledged and the same output is held constant, the various net elasticity estimates are consistent with net E–K complementarity.

III. A Reconciliation without the Separability Assumption

In the preceding discussion we have shown that engineering E–K substitutability, net E–K complementarity and the econometric results of the authors and G-G can be reconciled. Separability has played a prominent role in these reconciliations. We now show that the separability assumption, although useful for purposes of exposition and pedagogy, is not necessary for the reconciliation of engineering E–K substitutability with Hicks-Allen E–K complementarity. Thus our analytical findings are considerably generalized and strengthened.

It is well known that separability places restrictions on the Hicks-Allen partial elasticities of substitution σ_{ij} and price elasticities ϵ_{ij}.[11] Ryuzo Sato and Tetsunori Koizumi, among others, have defined the direct elasticity of substitution d_{ij} which in general differs from the Hicks-Allen elasticity measure whenever there are more than two inputs. In the present $Y = F(K, L, E, M)$ context, d_{KE} is defined as

$$(28) \quad d_{KE} = \frac{\partial \ln (K/E)}{\partial \ln (P_E/P_K)} \qquad Y, L, \text{ and } M \text{ fixed}$$

Note that the direct elasticity d_{KE} measures the price-induced change in the ratio of K to E conditional on the values of Y, L, and M, whereas the gross Hicks-Allen elasticity σ_{KE}^* measures the price-induced change in the level of K or E independently of Y, L, and M. Whether the recent engineering literature on E–K tradeoffs ignores independent inputs such as L or M (as in σ_{KE}^*) or merely assumes they are fixed (as in d_{KE}) is partly a matter in interpretation. The important role of $[(K, E), (L, M)]$ separability is that when it is imposed, the "conditional" and "independent" elasticity measures coincide, i.e., $d_{KE} = \sigma_{KE}^*$. However when this separability is not imposed, conditional and independent elasticity measures generally will not coincide. In particular, the d_{KE} elasticity which measures E–K substitutability conditional on L and M will be positive (this is assured by strict quasi concavity), even though K and E may be complementary inputs in the sense of Hicks-Allen. If one interprets engineering-economic studies as measuring E–K substitutability conditional on L and M, then resulting posi-

[11] For further discussion see Berndt and Christensen.

tive d_{KE} estimates can be completely consistent with negative, complementary values for the "unconditional" σ_{KE} estimates; this reconciliation does not depend on separability.

We now illustrate these points empirically.[12] When $[(K, E), (L, M)]$ separability is imposed on our data for total *U.S.* manufacturing, 1947–71, the estimates of σ^*_{KE} and d_{KE} coincide; the common value in 1965 is .243. However, when these separability restrictions (20) are not imposed, the conditional direct elasticity estimate d_{KE} in 1965 is .308, while the unconditional Hicks-Allen elasticity σ_{KE} is -3.193. Thus the positive conditional d_{KE} estimate occurs simultaneously with, and is perfectly consistent with, the negative σ_{KE} estimate. We conclude that a reconciliation of engineering E–K substitutability with Hicks-Allen complementarity does not depend on the convenient $Y = F[(K, E), (L, M)]$ separability assumption.

IV. Concluding Remarks

The purpose of this paper has been to provide an analytical and empirical interpretation of E–K complementarity consistent with basic microeconomics and with the process engineering evidence of E–K substitutability. Along the way we have developed the notion of utilized capital and have reconciled some of the seemingly disparate econometric findings. Above all, the analysis has emphasized to us that care must be taken in interpreting and properly comparing alternative elasticity measures.

A potentially promising research procedure suggested by our analytical framework is that if one is willing to assume utilized capital separability $Y = F[(K, E), L, M]$, then one can use either econometric or engineering estimates of gross E–K substitution elasticities in modeling industrial demand for energy. Engineering estimates might be preferable for longer-term forecasts involving new technologies, or for regions or countries with known technologies but unreliable economic data. These engineering estimates of parameters within the $K^* = f(K, E)$ submodel could then be integrated into an econometric model linking K^*, L, and M.[13]

Although we are not bound to the view that in all industries over any time period, E and K are complements, there appears to be a substantial and growing body of econometric evidence supporting the notion of Hicks-Allen E–K complementarity. Some of this research is empirical at a more disaggregated level,[14] while other analysis alters the definition of E–K complementarity.[15] In our view, however, empirical issues surrounding E–K complementarity remain unsettled, with a number of measurement and model specification problems worthy of particular attention in future research.

First, as of this date the available econometric evidence is based on data not including the post-1973 energy price increases. It would be useful to examine the robustness of the E–K complementarity findings with more recent data. Nonetheless, it is worth noting that E–K complementarity and E–L substitutability are consistent with the recent high-employment, low-investment recovery path of the *U.S.* economy. Irwin Kellner, a prominent bank economist, states this as follows:

> Finally, on the list of deterrents to capital spending, there is the significantly increased cost of building and operating new plants and equipment because of the higher price of energy of all types. The average economist may have forgotten his microeconomics, but the average businessman has not; he pays close attention to the relative cost of factors of production. And over the past three years it has become more expensive to increase capacity by adding machinery and equipment than it has by adding workers. [p. 3]

Second, there remains the concern, expressed vigorously by G-G, that E–K comple-

[12] Additional details are presented in our 1977 paper.

[13] Such a mixed process engineering-econometric modeling approach would be a natural extension of the integrated modeling efforts by Kenneth C. Hoffman and Jorgenson, and by Griffin.

[14] See, for example, Barry Field and Charles Grebenstein.

[15] Using a three-input *KLE* model of the aggregate *U.S.* economy, William Hogan shows that if L and P_K are fixed, recent increases in P_E are likely to result in reduced rates of capital formation.

mentarity estimates based on annual time-series data actually reflect short-run variations in capacity utilization, and that the "true" long-run relationship is one of E–K substitutability. For this reason G-G and Robert Pindyck prefer pooled cross-section time-series elasticity estimates to those based solely on time-series data. We have shown, however, that the G-G and Pindyck pooled cross-section time-series elasticity estimates must be interpreted carefully, and that in particular they are not inconsistent with our time-series result of E–K complementarity. Moreover, Fuss' pooled cross-section time-series data yield E–K complementarity. Hence the short-run, long-run E–K substitutability-complementarity issue does not seem to be simply one of pooled time-series data.[16] In our view, this empirical issue remains unresolved, for even if extremely reliable data were available, we still would need an economic model of the short-run adjustment process. At the present time, the econometric literature on dynamic adjustment processes relies largely on *ad hoc*, constant coefficient adjustment specifications, rather than on explicit dynamic optimization.[17]

Finally, a number of measurement problems remain. In their international pooled cross-section time-series studies, both G-G and Pindyck were unable to obtain data on variations in effective corporate and property tax rates among *OECD* countries and over time. Also, both studies computed the value of capital services as value-added minus the wage bill. This procedure has been criticized by, among others, Berndt (1976), for the resulting residual captures not only the return to capital equipment and structures, but also the returns to land, inventories, economic rent, working capital, indirect business taxes, and any errors in the measurement of value-added or the wage bill. Using 2-digit *U.S.* manufacturing cross-section *KLE* data for states in 1971, Field and Grebenstein obtain E–K substitutability when the return to capital is computed as value-added minus wage bill, but find E–K complementarity when the capital rental price measure refers only to plant equipment. Research on all these issues and in particular the inclusion of nonphysical inputs in the production function merits additional attention.

[16]In background work for his (1976b) paper, Edwin Kuh (1976a, Table 6c) finds that when the six "recession" or "excess capacity" years of 1949, 1954, 1958, 1961, 1970, and 1971 are dropped from our 1947–71 data set, E–K complementarity still results, albeit in a smaller absolute magnitude. Also, Berndt and Mohammed S. Khaled use our data and find that E–K complementarity is robust even when the assumptions of constant returns to scale and Hicks-neutral technical change are relaxed, and when flexible functional forms other than the translog cost function are employed.

[17]Research on this topic is presently under way; see Berndt, Fuss, and Leonard Waverman.

REFERENCES

E. R. Berndt, "Reconciling Alternative Estimates of the Elasticity of Substitution," *Rev. Econ. Statist.*, Feb. 1976, *58*, 59–68.

——, "Aggregate Energy, Efficiency and Productivity Measurement," *Ann. Rev. Energy*, Apr. 1978, *3*, 225–73.

—— **and L. R. Christensen**, "The Internal Structure of Functional Relationships: Separability, Substitution, and Aggregation," *Rev. Econ. Stud.*, July 1973, *40*, 403–10.

——, **M. A. Fuss, and L. Waverman**, "Dynamic Models of the Industrial Demand for Energy," rept. EA-500, Electric Power Res. Inst., Palo Alto, Nov. 1977.

—— **and D. W. Jorgenson**, "Production Structure," ch. 3 in Dale W. Jorgenson et al., eds., *U.S. Energy Resources and Economic Growth*, final report to the Ford Foundation Energy Policy Project, Washington 1973.

—— **and M. S. Khaled**, "Parametric Productivity Measurement and Choice Among Flexible Functional Forms," *J. Polit. Econ.*, forthcoming.

—— **and D. O. Wood**, "Engineering and Econometric Approaches to Industrial Energy Conservation and Capital Formation: A Reconciliation," work. paper no. MIT-EL 77-040WP, M.I.T. Energy Lab.; work. paper no. 16, Programme in Natural Resource Econ., Univ. British Columbia, Nov. 1977.

—— **and** ——, "Technology, Prices, and the Derived Demand for Energy," *Rev. Econ. Statist.*, Aug. 1975, *56*, 259–68.

C. **Blackorby, D. Primont, and R. R. Russell,** "Budgeting, Decentralization, and Aggregation," *Annals Econ. Soc. Measure.*, Winter 1975, *40*, 23–44.

T. G. **Cowing,** "Technical Change and Scale Economies in an Engineering Production Function: The Case of Steam Electric Power," *J. Ind. Econ.*, Dec. 1974, *23*, 135–52.

Joel Darmstadter et al., *How Industrial Societies Use Energy*, Baltimore 1977.

M. G. S. Denny and C. Pinto, "The Energy Crisis and the Demand for Labour," disc. paper no. 36, Res. Projects Group, Strategic Planning and Res. Div., Department of Manpower and Immigration, Ottawa, Aug. 1976.

——— **and M. Fuss,** "The Use of Approximation Analysis to Test for Separability and the Existence of Consistent Aggregates," *Amer. Econ. Rev.*, June 1977, *67*, 404–18.

B. C. Field and C. Grebenstein, "Capital-Energy Substitution in U.S. Manufacturing," mimeo., Univ. Massachusetts, Apr. 1978.

M. A. Fuss, "The Demand for Energy in Canadian Manufacturing: An Example of the Estimation of Production Structures with Many Inputs," *J. Econometrics*, Jan. 1977, *5*, 89–116.

J. M. Griffin, "Long-Run Production Modeling with Pseudo Data: Electric Power Generation," *Bell J. Econ.*, Spring 1977, *8*, 112–27.

——— **and P. R. Gregory,** "An Intercountry Translog Model of Energy Substitution Responses," *Amer. Econ. Rev.*, Dec. 1976, *66*, 845–57.

Elias P. Gyftopoulos, Lazaros J. Lazaridis, and Thomas F. Widmer, *Potential for Effective Use of Fuel in Industry*, Cambridge 1974.

——— **and T. F. Widmer,** "Energy Conservation and a Healthy Economy," *Tech. Rev.*, June 1977, *79*, 31–40.

R. Halvorsen and J. Ford, "Substitution Among Energy, Capital, and Labor Inputs in U.S. Manufacturing," in Robert S. Pindyck, ed., *Advances in the Economics of Energy and Resources: Structure of Energy Markets*, Vol. 1, Greenwich forthcoming.

J. A. Hausman, "Consumer Choice of Durables and Energy Demand," *Bell J. Econ.*, forthcoming.

E. Hirst, "Residential Energy Use Alternatives: 1976–2000," *Science*, Dec. 1976, *194*, 1247–52.

——— **and J. Carney,** "Residential Energy Use to the Year 2000: Conservation and Economics," Oak Ridge ORNL/CON-13, Sept. 1977.

E. Hirst et al., "An Improved Engineering-Economic Model of Residential Energy Use," Oak Ridge ORNL/CON-8, Apr. 1977.

K. C. Hoffman and D. W. Jorgenson, "Economic and Technological Models for Evaluation of Energy Policy," *Bell J. Econ.*, Autumn 1977, *8*, 444–66.

W. W. Hogan, "Capital Energy Complementarity in Aggregate Energy-Economic Analysis," *Bell J. Econ.*, forthcoming.

Leif Johansen, *Production Functions*, Amsterdam 1972.

I. Kellner, "Quarterly Business Conditions Analysis," Manufacturers Hanover Trust *Newsletter*, New York, Mar. 1977.

E. Kuh, (1976a) "Background Material for Research on Translog Cost of Production Function," unpublished paper, M.I.T., Feb. 1976.

———, (1976b) "Preliminary Observations on the Stability of the Translog Production Function," in *Proceedings of the Workshop Modeling the Interrelationships Between the Energy Sector and the General Economy*, Electric Power Res. Inst., spec. rept. no. 45, Palo Alto, July 1976.

J. R. Magnus, "Substitution Between Energy and Non-Energy Inputs in the Netherlands, 1950–1974," *Int. Econ. Rev.*, forthcoming.

M. Ohta, "Production Technologies of the U.S. Boiler and Turbogenerator Industries and Hedonic Price Indexes for their Products: A Cost Function Approach," *J. Polit. Econ.*, Feb. 1975, *83*, 1–26.

R. S. Pindyck, "Interfuel Substitution and the Industrial Demand for Energy: An International Comparison," work. paper no. MIT-EL 77-026WP, M.I.T. Energy Lab., Aug. 1977.

R. Sato and T. Koizumi, "On the Elasticities of Substitution and Complementarity," *Ox-*

ford Econ. Pap., Mar. 1973, *25*, 33–56.

L. Schipper and J. Darmstadter, "What is Energy Conservation?," rept. LBL-5919, Lawrence Berkeley Lab., Berkeley 1977.

R. H. Socolow, "The Coming Age of Conservation," *Ann. Rev. Energy*, 1977, *2*, 239–89.

P. Swaim and G. Friede, "Die Entwicklung des Energieverbrauchs der Bundes republik Deutschland und der Vereinigten Staaten von Amerika in Abhangigkeit von Preisen and Technologie," Karlsruhe: Inst. fur Angewandte Systemanalyse, June 1976.

Ford Foundation Energy Policy Project, *A Time to Choose: America's Energy Future*, Cambridge, Mass. 1974.

National Petroleum Council, *U.S. Energy Outlook: An Initial Appraisal, 1971–1985*, Washington 1971.

U.S. Federal Energy Administration, *Project Independence*, Washington, Nov. 1974.

U.S. National Academy of Sciences-National Research Council, *Report of the Demand/Conservation Panel*, Washington 1978.

U.S. Office of Emergency Preparedness, *The Potential for Energy Conservation: A Staff Study*, Executive Office of the President, Washington 1972.

Permanent Household Income and Consumption in Urban South America

By PHILIP MUSGROVE*

Milton Friedman's permanent income hypothesis (*PIH*) is considerably more complex and difficult to test than it appears at first sight, which is why certain of its postulates still cannot be accepted or rejected with certainty. One of the chief difficulties is that of separating the effects on consumption of permanent income, which is unobservable, and of various observable household characteristics which may be highly correlated with permanent income but may also be presumed to affect consumption directly. This difficulty can be overcome, provided a variable can be found which is associated with permanent income, but has no independent direct effect on the propensity to consume: if households are grouped by classes of that variable, the permanent consumption propensity or elasticity can be estimated from mean income and consumption in the groups. This technique requires the assumption of zero mean transitory income in each group, but does not depend on the postulate that transitory income is uncorrelated with consumption. For these reasons it cannot be applied to individual households.

This paper presents and estimates a model which separates the direct and indirect effects on consumption of a series of observable characteristics, using regression techniques on individual household observations. This makes it possible to estimate all the parameters of the model, both the consumption propensity or elasticity and the contribution to permanent income of each observable explanatory variable. The technique still requires that at least one of the explanatory variables affect consumption only through permanent income, but for all other variables both direct and indirect influences can be estimated, and the propensity to consume a given income can vary among households with different characteristics.

Another principal problem with the *PIH* is the assumption that transitory income and consumption are uncorrelated. Efforts to test this postulate can lead to very different results, ranging from agreement with the hypothesis to findings of high propensities to consume transitory income (see Robert Ferber, pp. 1306–09). This inconsistency, and continued controversy over the hypothesis, is due to the fact that permanent income cannot be estimated exactly, but can be known only with an error which is conceptually distinct from transitory income. Empirical work on the *PIH* has uniformly overlooked this fact, treating estimates of permanent income as exact and regarding the whole of the unexplained residual as transitory income.

The present model explicitly distinguishes the estimation error from the "error" of transitory income or consumption. The two transitory components can then be assumed to be uncorrelated, while the part of permanent income which is not explained is consumed exactly like the explained part. Whether residual income and consumption appear to be correlated or not then depends on the relative sizes of these two components, the unexplained permanent part and the true transitory part.

This technique has the further advantage that estimating the variance of the error of estimation makes it possible to tell how well a set of observable variables explain permanent income. This explanation may be better than

*The Brookings Institution. Robert Solow, Richard Eckaus, and Robert Hall supervised the dissertation (1974) on which this article is based. I am particularly grateful to Ross Williams and Robert Hall for help in specifying and estimating the model, and to Denisard Alves for a crucial suggestion. William Cline, Irwin Friend, Ricardo Ffrench-Davis, and Robert Ferber provided helpful comments on earlier versions. Subsequent discussions with Surjit Bhalla and Robert Lawrence have been most useful. I am also much indebted to the managing editor and to a referee for improvements in the final version.

that obtained for observed income, once the transitory variance is removed. It is also possible to estimate standard errors for the contributions of the observed variables to income and consumption.

The most implausible feature of the *PIH* is that it takes no account of needs. Friedman recognizes (pp. 121–23) that the propensity to consume may be higher for families with more children, but attributes this entirely to increased saving in the form of human capital accumulation, which makes other forms of saving less necessary for the household. Regarding children exclusively as a source of future income rather than of current needs may be defensible for wealthy families, but becomes implausible at low income. Friedman explicitly assumes that *any* reduction in current consumption can be compensated by a sufficient increase in future consumption, so that there is no time preference at any income level; he excludes from consideration "the possibility that the consumer unit will not live to engage in consumption in subsequent years" (fn. 8, p. 12). That possibility cannot be ignored by a majority of families in very poor countries, or by at least a few households even in rich countries; any theory applicable to their situation should allow for a threshold or subsistence level of consumption and therefore also of income. The present model does not incorporate a subsistence level because of the nonlinearity that would introduce, but the existence of permanent income levels low enough to preclude saving is regarded as influencing the estimation of average consumption propensities and the elasticity with respect to income.

The *PIH* has been little studied in less developed countries, where there is more reason to doubt the assumptions of a constant consumption propensity and of uncorrelated transitory components. Roger Betancourt has applied the hypothesis (1973) and the simpler "normal income" model (1971) to Chile; permanent income concepts have been used by Alan Kelley and Jeffrey Williamson for Indonesia; by R. Ramanathan and Firouz Vakil for India. A most thorough work for poor countries—and the only effort to take account of subsistence levels—is Surjit Bhalla's (1978, 1979) estimation for rural India. The present investigation covers households in seven cities of three South American countries.

I. The Permanent Income Model

Let us assume, following Friedman, that observed income Y and consumption C are decomposed into permanent (*) and transitory (**) components which are uncorrelated, and that the transitory elements have zero expected values:

$$Y = Y^* + Y^{**} \tag{1}$$

$$C = C^* + C^{**} \tag{2}$$

$$E(C^{**}) = E(Y^{**}) = 0 \tag{3}$$

$$E(C^*C^{**}) = E(Y^*Y^{**}) = 0 \tag{4}$$

$$E(C^{**}Y^{**}) = 0 \tag{5}$$

The only noninnocuous assumption is (5). Permanent income is assumed to be a function of m observable variables represented by a matrix X, with an error ϵ which also has zero expectation and is uncorrelated with either transitory component:[1]

$$Y^* = X\phi + \epsilon \tag{6}$$

where ϕ is an m x 1 vector;

$$E(\epsilon) = 0 \tag{7}$$

$$E(\epsilon C^{**}) = E(\epsilon Y^{**}) = 0 \tag{8}$$

Finally, permanent consumption is an exact function of permanent income and a set of r observable variables represented by a matrix

[1]The component ϵ is regarded here as just as much a part of permanent income as the explained component It is an "error" only for the analyst: the consumer knows exactly what his permanent income is. A more complex, and more plausible, interpretation due to Michael Landsberger and Arthur Okun, is that ϵ contains in addition to the estimation error, an uncertainty on the part of the consumer. Permanent income is known only with a margin of error, and therefore so is transitory income. Small changes in current income are treated as permanent, while large changes (windfalls) are regarded as transitory. This refinement cannot be incorporated in the present model, since it requires that C^* vary nonlinearly with ϵ.

H:

$$C^* = H\theta + \mu Y^* \tag{9}$$

where θ is an $r \times 1$ vector.

Expressions (1) through (9) can describe the linear form of the *PIH*, with μ being the *MPC*, and the proportionality hypothesis requiring that $H\theta = 0$; or the logarithmic form, in which the elasticity μ must be unity if $APC = MPC$, and the variables H exert constant proportional effects on consumption. Only the logarithmic version is considered here. I also assume that H is a subset of X with $r < m$, so that some variables are allowed to affect consumption both directly (θ_i) and indirectly ($\mu\phi_i$), while the remaining $m - r$ variables, denoted Z, affect consumption only via income. In order to estimate Y^*, there must be at least one variable in Z, requiring $r < m$.

Substitution of (6) and (9) into (1) and (2) yields the reduced form:

$$\begin{aligned} (10)\quad Y &= X\phi + \epsilon + Y^{**} = X\phi + V_Y \\ &= H\phi_H + Z\phi_Z + V_Y \\ C &= H\theta + \mu X\phi + \mu\epsilon + C^{**} \\ &= H\theta + \mu X\phi + V_C \\ &= H(\theta + \mu\phi_H) + \mu Z\phi_Z + V_C \end{aligned}$$

and (5) and (8) mean that the covariance matrix of the compound errors V_Y, V_C can be interpreted as

$$\begin{aligned} (11)\quad V_{YY} &= var(V_Y) = var(\epsilon) \\ &\quad + var(Y^{**}) \\ V_{CC} &= var(V_C) = \mu^2 var(\epsilon) \\ &\quad + var(C^{**}) \\ V_{CY} &= cov(V_C, V_Y) = \mu\, var(\epsilon) \end{aligned}$$

Note that assumption (5) is necessary for identification: this allows the residual variation V_{YY} or V_{CC} to be decomposed into a part which is transitory in Friedman's zero correlation sense and a part which is common to both income and consumption; which component appears to predominate depends on μ and on the relative magnitudes of V_{YY}, V_{CC}, and V_{CY}.

The requirement that the transitory variances be nonnegative confines μ to the interval defined by V_{CY}/V_{YY} and V_{CC}/V_{CY}. Nonnegativity of $var(\varepsilon)$ sets lower bounds to the transitory variances, provided $V_{CY} > 0$. Finally, the error variances can be used to define R^2 statistics.

$$\begin{aligned} (12)\quad R^2_{Y^*} &= \frac{var(\hat{Y}^*)}{var(Y^*)} \\ &= \frac{var(Y) - var(Y^{**}) - var(\epsilon)}{var(Y) - var(Y^{**})} \\ R^2_{C^*} &= \frac{var(\hat{C}^*)}{var(C^*)} \\ &= \frac{var(C) - var(C^{**}) - \mu^2 var(\epsilon)}{var(C) - var(C^{**})} \end{aligned}$$

which show how well H and Z explain the permanent components. These statistics are not derived from regressions, since they refer to the unobservable variables; but $var(Y)$ and $var(C)$ can be observed and the three error variances can be recovered from (11), given estimates of μ and of V_{YY}, V_{CC}, and V_{CY}. The derivation of the $m + r + 1$ parameters of the model (m elements of ϕ, r of θ, and μ) is described in detail in my thesis and is summarized in the Appendix. All the parameters except μ can be estimated by ordinary least squares (*OLS*) regression after transformation of the variables.

II. Results for Urban South America

The empirical results are based on household budget data collected in 1967–69 from four principal cities in Colombia, two in Ecuador, and one in Peru. Some of the parameter estimates have been reported elsewhere (see the author, 1978b, which also describes the data in detail and reports other analyses, and 1978a), but most are presented here for the first time.

The household surveys, which proceeded in four waves of interviews during a period of 12–15 months, form part of the study of income and expenditure of the ECIEL Program of Joint Studies of Latin American

Economic Integration:[2] the data are conceptually quite comparable and were subjected to a uniform series of logical and statistical tests. In Colombia and in Ecuador, some households were interviewed between two and four times, but the period covered by the panel data is too short to provide any evidence about the length of the consumer horizon and so all observations are treated independently.[3] The model was estimated once for each country: individual noncapital cities (Barranquilla, Cali, and Medellín in Colombia; Guayaquil in Ecuador) were distinguished by dummy variables from the capitals (Bogotá and Quito). These variables were included in H to allow intercity variation in consumption propensities as well as in income levels.

The term Y is defined as disposable income—total income less direct taxes and Social Security payments—less income which can be classified *ex ante* as transitory and is very infrequently received: retirement bonuses, inheritances, lottery winnings, and gifts. Seasonal bonuses, which are anticipated and therefore likely to be regarded as permanent, are included. Income in kind, including imputed rent on owned dwellings, is also part of income. C is defined as discretionary expenditure, excluding tax and Social Security payments, and including half of expenditure on durable goods.[4]

The vector Z consists of fifteen dummy variables. Three refer to classes of occupation of the household head and twelve describe combinations of age and formal schooling, also for the head. Four age groups and four educational levels are distinguished, but there is no interaction with age for the group of uneducated household heads. The use of age as a determinant of permanent income means that the household's horizon is assumed to be, at most, the length of an age bracket (usually fifteen years): the spirit of the assumptions is closer to Friedman than to the life cycle hypothesis (*LCH*) of Franco Modigliani and Albert Ando.

The crucial assumption in the definition of Z and H is that age, education, and occupation have no direct effect on consumption, apart from their indirect influence through permanent income. This assumption is most defensible for age, since life cycle effects in the propensity to consume are introduced by including dummy variables for six life cycle stages in H. There is then little reason to expect an additional "pure" age effect. There is also no reason for a pure occupational effect on permanent consumption, apart from the possession of human or physical capital associated with a particular job or profession. (People in different occupations may have very different *apparent* consumption propensities, as Friedman has noted in comparing farm and nonfarm groups, but that reflects different variances of transitory income rather than different permanent propensities. Thus the limitation of the present model is primarily its assumption of homoscedastic transitory components.)

The assumption of no direct consumption effect is weakest for education, since schooling is a form of wealth whose possession might lead people to vary their saving in other forms. This assumption could best be tested by including in the model household wealth in all forms, but the data are incomplete and unsatisfactory for some assets and it would be necessary, even with better data, to value human capital for this comparison. A simpler test is to include in H the share of total income received from capital, a continuous variable. This variable affects both income and the propensity to consume, so it provides a (weak) test of Friedman's hypothesis that

[2]The institutions whose data are used are the Centro de Estudios sobre Desarrollo Económico of the Universidad de los Andes, Bogotá, Colombia; the Instituto Nacional de Estadística y Censos, Quito, Ecuador; and the Centro de Investigaciones Sociológicas, Económicas, Políticas y Antropológicas of the Pontificia Universidad Católica del Perú, Lima. I am indebted to my colleagues Guillermo Perry, Jorge Yépez, and Adolfo Figueroa for collaboration in this study and for permission to use their respective institutions' data. There are 2,943 observations for Colombia, 1,970 for Ecuador, and 1,328 for Peru.

[3]A further reason for not combining observations of the same household is the assumption of homoscedastic errors: families might be expected to show different transitory variances as the interval of observation varies.

[4]The data do not provide any estimate of consumption (depreciation) of durables. Except for life cycle effects, all the parameters are quite insensitive to the share of durables included in C.

the propensity is related to the share of the household's wealth which is in nonhuman form: for a given permanent income, a higher share should mean a lower need to accumulate assets, particularly liquid assets, and a higher rate of consumption. The H variables also include two descriptions of household employment: a dummy variable indicating the head is not employed (usually but not always retired) and another indicating that more than one member is employed. Altogether, H contains nine variables, apart from cities; it also includes a constant term, which absorbs the effects of one age-education class, one occupation, one life cycle stage, and the capital cities.[5]

A. *Parameter Estimates*

Estimates of μ and its standard error are shown in Table 1, together with lower and upper limits for μ, and a current elasticity estimated from the relation $C = \mu_0 + \mu_1 Y$, together with its standard error σ_1. For each country, μ is clearly distinguishable from the short-term elasticity μ_1, and also from unity. On the assumption of a constant elasticity, the proportionality hypothesis is rejected. The results instead support Thomas Mayer's "standard income" hypothesis, according to which $0 < \mu_1 < \mu < 1$. Of course, a constant nonunitary elasticity is theoretically unsatisfactory if $H\theta > 0$, since at low enough income it leads to $C^* > Y^*$.

The estimates of μ for Colombia and Ecuador are indistinguishable, but that for Peru is clearly lower. Since incomes are comparable in these countries, there is no particular reason why μ should differ.[6]

TABLE 1—ESTIMATES OF μ (PERMANENT CONSUMPTION ELASTICITY) AND TRUE STANDARD ERRORS

	Colombia	Ecuador	Peru
μ (Elasticity)	0.881	0.896	0.776
σ_μ (True Standard Error)	0.015	0.013	0.023
t-Statistic for Test of $\mu = 1$	7.933	8.00	9.739
Limiting Values of μ			
Minimum	0.756	0.730	0.413
Maximum	1.029	1.100	1.405
Current Consumption Elasticity (μ_1)	0.813	0.812	0.555
OLS Standard Error (σ_1)	0.007	0.009	0.016
t-Statistic for $\mu = \mu_1$	4.108	5.313	7.888
t-Statistics for Tests of Equality of μ Across Countries	Colombia	0.756	3.824
	Ecuador		4.542

However, the Peruvian sol was devalued by 40 percent in September 1967, just five months before the survey began, and inflation was subsequently much more rapid in Peru than in Colombia or Ecuador. It is therefore to be expected that there was much more transitory income variation in Peru. This is suggested by the very low current elasticity μ_1 and the relatively high ratio μ/μ_1 for that country (1.4 in Peru, vs. 1.1 in Colombia and Ecuador). The large transitory changes may also have upset the income-determining relation, so that ϕ was unstable during the year. In that case μ would also be biased; under more normal circumstances it might be about 0.9 rather than below 0.8.

Estimates of ϕ_Z are presented in Table 2, together with the standard errors of the regression used to estimate it. These show all the variables to be significant in Colombia, and all but two in Ecuador and Peru. Because these estimates understate the true standard error, two F-tests were applied to the Peruvian results, which come from the smallest sample and show the largest σ_μ. The "occupational" variables were tested as a group and

[5]For Colombia only, H includes a dummy variable indicating a secondary consuming unit, which shares part of its expenses (chiefly rent) with a primary household, but maintains a separate budget for other expenditures. Such units constitute only 5 percent of Colombian households; the corresponding parameters are not reported here.

[6]Median consumption per person in households is just over $500 per year in Bogotá and Lima, between $400 and $450 per year in Barranquilla, Cali, Quito, and Guayaquil, and just under $350 per year in Medellín. Values in local currency are converted to *U.S.* dollars by converting to Venezuelan bolívars of equal purchasing power in May 1968; adjusting for inflation between the date and the survey period; and converting to dollars by an estimated parity rate (4.37) for the bolívar. (See my book, pp. 27–31; 185–87.)

TABLE 2—ESTIMATES OF ϕ_Z (INCOME COEFFICIENTS OF AGE, EDUCATION, AND OCCUPATION) AND *OLS* STANDARD ERRORS

	Colombia	Ecuador	Peru
Age/Education Classes[a]			
(All Ages) No Education	−0.207	−0.294	−0.112
	(0.055)	(0.096)	(0.163)
35–49 Primary	0.074	0.097	0.118
	(0.037)	(0.051)	(0.074)
50–64 Primary	0.111	0.157	0.323
	(0.047)	(0.062)	(0.084)
65 or over Primary	0.305	0.083	0.224
	(0.081)	(0.093)	(0.120)
12–34 Secondary	0.532	0.483	0.413
	(0.043)	(0.061)	(0.083)
35–49 Secondary	0.625	0.896	0.706
	(0.043)	(0.050)	(0.079)
50–64 Secondary	0.719	0.964	0.658
	(0.061)	(0.075)	(0.095)
65 or over Secondary	1.070	0.824	0.788
	(0.119)	(0.137)	(0.151)
12–34 University	1.160	0.946	0.756
	(0.079)	(0.100)	(0.104)
35–49 University	1.344	1.238	1.234
	(0.083)	(0.103)	(0.094)
50–64 University	1.518	1.294	1.113
	(0.104)	(0.139)	(0.123)
65 or over University	1.820	1.590	1.523
	(0.257)	(0.191)	(0.188)
Occupation Classes[b]			
Professional, Technical, and Managerial	0.294	0.270	0.435
	(0.046)	(0.066)	(0.063)
Blue Collar (Laborers and Artisans)	−0.147	−0.187	−0.109
	(0.028)	(0.042)	(0.050)
Other Occupations	−0.237	−0.304	0.015
	(0.035)	(0.047)	(0.062)
R^2 Statistic for Regression	0.399	0.360	0.344

[a]Coefficient of 12–34 Primary is restricted to zero.
[b]Coefficient of white-collar occupations is restricted to zero.

found to be highly significant. The variable "35–49 Primary" was also tested, yielding a t-statistic of 2.36 in place of the *OLS* value of 2.55. Without repeating the F-test for every variable, it cannot be said what all the true errors are, but it seems safe to conclude that they are only about 10 percent larger than the *OLS* estimates. The results for Peru can be applied to the other countries, where μ and ϕ_Z are estimated more precisely.

The same pattern of age-education effects is found, with minor exceptions, in all three countries: permanent income rises with age and with education, the increase with age being steeper, the higher the level of schooling. Beyond age 64, income declines, except for the university educated. At lower schooling levels, age differences are not always statistically significant, but educational differences usually are.

Estimates of ϕ_H are presented in Table 3, again with *OLS* standard errors. No F-tests were made of these coefficients, because each element of ϕ_H depends on all the coefficients of ϕ_Z and their associated errors. I assume

TABLE 3—ESTIMATES OF ϕ_H (INCOME COEFFICIENTS OF LIFE CYCLE, EMPLOYMENT, CAPITAL INCOME SHARE, AND CITY) AND *OLS* STANDARD ERRORS

	Colombia	Ecuador	Peru
Constant *log* (National Currency)	5.932	7.944	6.874
	(0.026)	(0.035)	(0.040)
Constant *log* (*U.S.* $ per Year)	7.265	6.807	7.107
Life Cycle Stage[a]			
Unmarried	−0.182	−0.297	−0.086
	(0.040)	(0.052)	(0.107)
Married, No Children	−0.084	−0.086	−0.015
	(0.065)	(0.097)	(0.128)
Oldest Child Aged 8–18	0.166	0.005	0.040
	(0.026)	(0.035)	(0.046)
Oldest Child Aged 19 or More	0.145	−0.107	0.061
	(0.050)	(0.059)	(0.069)
Retired	0.112	−0.414	−0.256
	(0.079)	(0.107)	(0.166)
Head Not Employed	−0.058	−0.004	−0.091
	(0.042)	(0.054)	(0.066)
More Than One Member Employed	0.195	0.436	0.356
	(0.030)	(0.037)	(0.040)
Share of Income from Capital	0.007	0.018	0.006
	(0.001)	(0.001)	(0.001)
City[b]			
Barranquilla	−0.048		
	(0.035)		
Cali	−0.094		
	(0.032)		
Medellin	−0.063		
	(0.029)		
Guayaquil		0.298	
		(0.030)	
R^2 Statistic for Regression	0.121	0.269	0.074

[a]Coefficient for married, young children is restricted to zero.
[b]Coefficient of capital city is restricted to zero.

instead that the results for ϕ_Z apply here; true standard errors probably do not exceed the *OLS* estimates by more than 10 percent. (Only the coefficient for "Medellín" would cease to be significant if the t-test limit were raised from 1.96 to 2.5, implying a 28 percent increase in the standard error.)

The presence of additional income earners always contributes substantially to income. The effect is least in Colombia because the sample excludes supplementary members of the household, who are always adults and who share only part of the family budget; the income provided by spouses and children is likely to be smaller and more transitory. The nonemployment of the head, however, has no significant effect on Y^*, after allowing for age, education, and occupation. Most unemployment is probably transitory and, as Albert Berry found for Colombia, is not associated with low incomes. Not surprisingly, capital income always has a significant impact: if 10 percent of income is derived from capital, permanent income is 6–18 percent higher.[7]

Permanent income initially rises in the life cycle, up to the point where the family includes adolescent children, and declines, sometimes quite sharply, in retirement. These results are consistent with the assumption of a horizon considerably shorter than the lifetime of the household. Neither very young nor very old people appear to include in their notion of permanent income the higher incomes they will (have) recieve(d) in their peak earning years. Since age and the number of people employed have been taken into account, however, the standard errors are relatively large; in Peru, none of the life cycle variables appears significant.

The interpretation of the vector θ is most dependent on assumptions about the consumer's horizon. Parameter estimates and errors are presented in Table 4. An F-test of the "Retired" variable for Peru suggests the true standard errors exceed the *OLS* estimates by less than 5 percent. The propensity to consume rises when the head is not employed: for Colombia, θ_i and $-\mu\phi_{Hi}$ are about equal, so that nonemployment has no *net* effect on consumption, but permanent consumption appears to fall sharply in Peru and to rise in Ecuador. Since retirement is considered separately, this change in C^* is not primarily a life cycle effect. Consumption tends to decline relative to income when more than one member is employed. This suggests that additional members may seek employment specifically so the family can save more, perhaps in the form of durables.

The share of income from capital appears not to affect the consumption propensity in Colombia and Ecuador, but to raise it in Peru. However, it is difficult to interpret these results in terms of Friedman's hypothesis

TABLE 4—ESTIMATES OF θ (COEFFICIENTS AFFECTING THE PROPENSITY TO CONSUME) AND *OLS* STANDARD ERRORS

	Colombia	Ecuador	Peru
Constant	0.862	0.985	1.706
	(0.012)	(0.019)	(0.029)
Life Cycle Stage			
Unmarried	−0.064	0.058	−0.062
	(0.019)	(0.028)	(0.080)
Married, No Children	−0.052	0.007	0.139
	(0.031)	(0.053)	(0.096)
Oldest Child 8–18	0.006	0.070	0.037
	(0.012)	(0.019)	(0.034)
Oldest Child 19 or More	−0.031	0.068	−0.015
	(0.024)	(0.032)	(0.051)
Retired	−0.071	0.012	−0.005
	(0.037)	(0.058)	(0.124)
Head Not Employed	0.067	0.180	0.039
	(0.020)	(0.029)	(0.049)
More than One Member Employed	−0.039	−0.003	−0.105
	(0.014)	(0.017)	(0.030)
Share of Income from Capital	−0.0004	−0.001	0.005
	(0.0003)	(0.001)	(0.001)
City			
Barranquilla	0.072		
	(0.017)		
Cali	−0.011		
	(0.015)		
Medellín	−0.133		
	(0.014)		
Guayaquil		−0.108	
		(0.016)	
R^2 Statistic for Regression	0.064	0.075	0.041

[7]Since the independent variable is a share of current income, these coefficients are not elasticities of permanent income with respect to capital income. There is no implication that a change in capital income affects Y^* differently than a change in any other kind of income.

about the composition of wealth, because for most households, the bulk of capital income is imputed rent on owned dwellings, a relatively illiquid form of wealth and a measure which takes no account of the amount of equity the family owns in the dwelling.[8] Imputed rent is a fraction of market value, and so is independent of the outstanding mortgage. The positive effect of this variable in Peru may arise because capital income included less transitory variation than labor income—that is, a high estimate of θ_i may simply reflect the presumed downward bias in μ discussed earlier.

Consumption tends to be lower, relative to income, in noncapital cities, and particularly in major industrial centers (Medellín and Guayaquil). It is not clear whether this frugal behavior characterizes the entire population, including the working class and the poor, or whether it reflects the fact that the rich in such cities are more likely to be entrepreneurs with high propensities to invest, whereas the rich in the capital cities include more landowners and other people with low saving propensities. (Permanent income differs significantly between cities in Ecuador, but not in Colombia.)

In Colombia and Ecuador, the propensity to consume is highest for families with adolescent children (aged 8 to 18), which are likely to be the largest households and thus to have the greatest needs. In Peru, the propensity is even higher for young married couples with no children: I attribute this to high spending on durables, half of which is included in consumption. Durables account for 8 percent of total expenditure in Peru, vs. 3 percent in Colombia and 4 percent in Ecuador, and the high share appears to be due to rapid and accelerated inflation at the time of the survey, which made durables more attractive than other forms of saving, particularly to middle income families with no access to financial assets paying protected real returns. (This argument is developed further in my paper with Howard Howe.)

With few exceptions, the life cycle pattern is for the propensity to rise as the family is built up and needs become greater, and then to fall as children leave home. The total variation over the cycle is—except for the acquisition of durables by young households in Peru—about 7 percent of consumption in Colombia and Ecuador, and 10 percent in Peru. The most surprising result is the low spending propensity among retired households, who might be expected to live partly off their savings and so spend a larger share of income. Apart from the treatment of durables (of which older families buy very little), failure to observe this decumulation among the elderly probably reflects the fact that under variable and sometimes high inflation, and with extremely imperfect capital markets, it is impossible to accumulate for retirement with any certainty. It may also matter that pensions, Social Security, and other transfers reach a much smaller share of the population than in high-income countries, making retirement incomes not only lower but more variable.

The vectors ϕ_Z, ϕ_H, and θ are estimated by regressions which yield R^2 statistics, also given in Tables 2, 3, and 4. These show that ϕ_Z is always easier to estimate than ϕ_H, which in turn is estimated more easily than θ: the more an effect refers to consumption rather than income, the harder it is to isolate. The very low value of R^2 (ϕ_H) for Peru suggests that it may be the H variables, rather than those in Z, whose relation to Y^* was upset by transitory factors; R^2 (ϕ_Z) is not very different among the three countries.

B. *Propensities to Consume*

In the linear form of the *PIH*, the *APC* equals the *MPC* and, assuming zero expected transitory errors, the propensity can be estimated from the observed variables. This is not true under the logarithmic form analyzed here. The observed *APC* is $k = E(c)/E(y)$, where c and y are the variables rather than their logarithms, whereas the permanent *APC*

[8]A better test of the Friedman hypothesis is provided by Jean Crockett and Irwin Friend, who use the same data for Colombia to estimate the propensity to consume normal income. They include the variable Y_l/Y_d where Y_d corresponds closely to Y, and Y_l is dividend and interest income only. This financial wealth variable raises consumption significantly.

is $k^* = E(c^*)/E(y^*)$. The estimate of k^* can be found from the imputed values $\hat{c}^* = \exp(H\hat{\theta} + \hat{\mu}X\hat{\phi})$ and $\hat{y}^* = \exp(X\hat{\phi})$, assuming the error ϵ to be normally distributed (see J. Aitchison and J. A. C. Brown, p. 111), as

$$\frac{E(\hat{c}^*)\exp(0.5\,\mu^2\,var(\epsilon))}{E(\hat{y}^*)\exp(0.5\,var(\epsilon))}$$

This is less than $E(\hat{c}^*)/E(\hat{y}^*)$ when $\mu < 1$, but it is not evident whether k^* will be more or less than k. Estimates for the three countries are given in Table 5. In every case, k^* is a lower, and more plausible, estimate than k: this means that average transitory income in arithmetic term was negative in all three countries, raising the observed *APC*. The amount of negative transitory income was largest in Peru and smallest in Colombia.

C. *Contribution of Income and Consumption to Permanent Income*

Both ϕ_Z and θ are estimated from regressions whose dependent variables are linear combinations of C and Y; the coefficients in ϕ_Z are respectively α_c and α_y, and in θ they are 1 and μ. Since $\mu < 1$ for all three countries, C contributes more than Y to the estimation of both vectors. The values of the relative importance of C in each case are also given in Table 5. For each country, the shares are very similar; intercountry differences reflect mostly differences in μ, which in turn are hypothesized to reflect differences in transitory variation and in the stability of the relations indicated by ϕ and θ.

TABLE 5—ESTIMATES OF THE AVERAGE CONSUMPTION PROPENSITY AND OF THE RELATIVE IMPORTANCE OF CONSUMPTION IN ESTIMATING THE VECTORS ϕ_Z, θ

	Colombia	Ecuador	Peru
Average Consumption Propensities			
Observed (k)	1.0109	1.0040	1.0035
Permanent			
Calculated:			
$E(\hat{c}^*)/E(\hat{y}^*)$	1.0321	0.9707	0.9752
Assuming Lognormal Errors (k^*)	0.9966	0.9391	0.9240
Relative Weight of Consumption in Estimating			
$\phi_Z : \alpha_c/(\alpha_c + \alpha_y)$	0.5273	0.5209	0.5825
$\theta : 1/(1 + \mu)$	0.5316	0.5272	0.5631

It is clear from these results that observed consumption is a better proxy for permanent income than is observed income, and—not surprisingly—a better proxy for permanent consumption. It is also evident, however, that observed income contributes much information toward estimating permanent income, so a better measure will combine C and Y. An even better proxy can be formed if C and Y are replaced by instrumental variables, obtained by regressing them on the hypothesized determinants Z of permanent income. In the best of circumstances, it might be possible to find a set of variables $X = (H, Z)$ such that H is uncorrelated with Z and Y, while H is correlated with C, and Z with both Y and C. Then permanent income could be estimated simply by $Z(Z'Z)^{-1}Z'(C + Y)$. The assumption that transitory effects are uncorrelated makes a linear combination of C and Y better than either variable alone; transitory components are more likely to cancel out. The use of instrumental variables further reduces the transitory components, provided they are uncorrelated with Z.

D. *The Decompositoin of Observed Income and Consumption*

From the reduced form (10), the variance of income can be separated into an explained permanent component, an unexplained or residual permanent component, and a transitory component; and similarly for consumption. The resulting estimates are shown in Table 6. In Colombia, transitory variation amounts to only 7.8 percent of the total variance for both income and consumption, while in Ecuador, the shares are respectively 8.8 and 8.7 percent. As expected, transitory variation was much more important in Peru: 31.5 percent of income variance and 27.0 percent for consumption. In all three countries, the error of estimation ϵ has a larger variance than the transitory components (much larger in Colombia and Ecuador), so that failure to consider it could lead to large

TABLE 6—DECOMPOSITION OF VARIANCE, OBSERVED INCOME AND CONSUMPTION

	Colombia	Ecuador	Peru
Variance of			
Income (Y)	0.666	0.884	0.753
Transitory Income (Y^{**})	0.052	0.078	0.237
Permanent Income (Y^{*})	0.614	0.806	0.515
Estimation Error (ϵ)	0.313	0.336	0.271
Explained Component ($X\phi$)	0.301	0.470	0.245
Age Variation	0.007	0.018	0.021
Consumption (C)	0.525	0.714	0.489
Transitory Consumption (C^{**})	0.041	0.062	0.132
Permanent Consumption (C^{*})	0.484	0.652	0.357
Permanent Error ($\mu\epsilon$)	0.243	0.270	0.163
Explained Component ($H\theta + \mu X\phi$)	0.241	0.382	0.194
R^2 Statistics for			
Y^*	0.491	0.583	0.475
C^*	0.497	0.587	0.543
Y	0.452	0.526	0.326
C	0.458	0.530	0.397

overestimates of the importance of transitory variation.

The finding that under normal conditions only a small share of income variance is due to transitory factors is extremely important, because it says that nearly all the income concentration observed in these countries is permanent, even when the data refer to a relatively short interval. The "problem" of income inequality is not therefore an illusion created by short-run movements with no welfare impact. However, this result is due in considerable measure to the age-dependent definition of permanent income. As Vladimir Stoikov, Morton Paglin, and Simon Kuznets have pointed out for the United States, lifetime incomes differ less than observed incomes, because much income variation is due to age differences among people on similar age-income profiles. This may also be true in Latin America, at least for highly educated people. At low levels of schooling, however, there is little variation with age, so that much of the permanent variance is due to schooling differences, and lifetime incomes are still very unequally distributed.

It is not possible to separate age and lifetime effects entirely in this model, because there may be interactions between age and variables other than education, which are not taken into account. (In particular, age may interact with capital income, a high capital share having a different effect for young than for old households.) Assuming there are no such interactions, a between-age effect can be defined by summing the between-age variations at each level of schooling, and the rest of permanent income variance can be assumed to reflect lifetime differences.[9] The variance attributable to age differences alone is extremely small, as indicated in Table 6: it amounts to only 1.2 percent of total permanent income variation in Colombia, 2.3 percent in Ecuador, and 4.1 percent in Peru. These estimates may understate the age effect because of omitted interactions, and because no age variation is considered among the uneducated. It does not follow that the amount of variation which is important for welfare is overstated, because age variation is not really "compensated" over a household's lifetime. The extreme imperfection of capital markets makes it impossible for consumers to rearrange their lifetime incomes to yield the best consumption path, and the problem is worsened by changes in household size as children are added, which make the time path of consumption per person still more variable.

The finding of a small variance for transitory income also does not necessarily mean that welfare is very little affected. For consumers who cannot borrow and lend freely and at comparable interest rates, any transitory variation at all can impose a considerable welfare loss in the form of interest paid or foregone, or in the form of unwanted variations in consumption. Moreover, transitory income may not be homoscedastic; in fact, its variance may be expected to be highest at the extremes of the (permanent) income distribution, and lowest for the middle incomes asso-

[9]If ϕ_{EA} is the income coefficient for an age-education class, and N_{EA} is the share of households in that class, then $\phi_E = \Sigma_A N_{EA}\phi_{EA}/\Sigma_A N_{EA}$ is the mean coefficient for that education class, and the class variance is $\Sigma_A N_{EA}(\phi_{EA} - \phi_E)^2/\Sigma_A N_{EA}$. Since $\Sigma_E \Sigma_A N_{EA} = 1$, the total between-age variance is simply $\Sigma_E \Sigma_A N_{EA}(\phi_{EA} - \phi_E)^2$.

ciated with regular salaried employment.[10] An overall variation of 8 percent of income may conceal variations of 20–30 percent among the poor, who would be most affected by such irregularity. Transitory variation might also differ appreciably across life cycle groups, particularly at the extremes of the cycle.

Table 6 shows the statistics $R^2_{Y^*}$ and $R^2_{C^*}$ as defined earlier in (12) and the corresponding *OLS* statistics R^2_Y and R^2_C associated with regressing Y and C on X. In Colombia and in Ecuador, $R^2_{Y^*}$ and $R^2_{C^*}$ are very similar, being about 0.5 in the first country and almost 0.6 in the second. Comparable statistics are obtained for Peru, except that permanent consumption is explained appreciably better than permanent income. The R^2 statistic for the unobservable variable always exceeds that for the corresponding observable variable. It may be noted that the much higher transitory variation in Peru does not seriously affect the ability to explain the variation atttributed to permanent factors.

III. Summary

This investigation yields three principal substantive findings, equally applicable to all the countries studied. First, pure transitory variation in income—that part which is uncorrelated with consumption—is quite small, unless the entire economy is subjected to a net transitory shock. The bulk of residual or unexplained income is best regarded as an error of estimation which is correlated with consumption, rather than as a transitory component. Nearly all income variation then is permanent and corresponds to long-lasting differences in welfare.

Second, permanent income can be estimated satisfactorily from a small number of variables. At least in urban Latin America, education is by far the most powerful variable in explaining income differences. For the well educated, income also varies substantially with age. There is little further effect on income from life cycle variables, although these may significantly affect the propensity to consume.

Third, the average elasticity of consumption with respect to permanent income is clearly less than one, and also clearly more than the observed elasticity in the short run. Accepting all of Friedman's assumptions except that of proportionality leads to an unequivocal rejection of the proportionality hypothesis, with an elasticity of about 0.9. This average value is compatible with a unitary elasticity both at very low incomes (where saving is impossible because of subsistence needs) and at high incomes, where the proportionality assumption becomes plausible. A nonunitary elasticity such as that found here probably characterizes a transition region of saving behavior in countries where many families live close to subsistence but average income is high enough to permit some households to save systematically.

[10]These suggestions are supported by the finding that wage and salary incomes are concentrated in the middle two quartiles of the distribution, while self-employment incomes—which probably include more transitory variation—are found predominantly in the highest and lowest quartiles.

Appendix—Estimation of Parameters and Standard Errors

The model is estimated following a procedure developed by Arnold Zellner on the assumption $\epsilon = 0$, and extended by Arthur Goldberger to the case of positive $var(\epsilon)$; Karl Jöreskog includes it as one of several related structural equation models. The procedure is identical if $var(\epsilon) = 0$, but (5) is relaxed, so that C^{**} and Y^{**} may be correlated. This tests the zero correlation assumption, but only on the implausible condition that Y^* can be estimated without error. C. L. F. Attfield uses this approach to test the propensity to consume transitory income, which is determined by a subset of the explanatory variables; there is no variation in the permanent propensity in this model. The zero correlation assumption can also be tested, while retaining the assumption of $var(\epsilon) > 0$, in a model described in my 1977 article, where C^* is the sum of consumption in several categories, each linearly related to Y^*.

Let A be the matrix $[Y, C]$, V the matrix $[V_Y, V_C]$, and π the matrix of true parameter

values

$$\pi = \begin{bmatrix} \phi_H & \theta + \mu\phi_H \\ \phi_Z & \mu\phi_Z \end{bmatrix}$$

Then the reduced form (10) can be written $A = X\pi + V$. Let P be the *OLS* estimator of π, obtained by regressing Y and C on X, $P = (X'X)^{-1}X'A$ and define the matrices $W = (A - X\pi)'(A - X\pi)$ and $S = (A - XP)'(A - XP)$, the *OLS* estimator of V. The covariance matrix of the true errors is $\Omega = E[V'V]$.

Assuming the errors to be normally distributed, the *log* likelihood function for the reduced form can be written $L(\Omega, \pi) = log\, det(\Omega^{-1}) - tr(\Omega^{-1}W)$. Maximizing this is equivalent to choosing π to minimize $Q = det\,(S^{-1}W)$ since $det\,(S^{-1}W) = det\,(S^{-1})\, det(W)$, and $det(S^{-1})$ is a positive constant. However, the same choice of π minimizes the sum and the product of the characteristic roots of $S^{-1}W$, as is shown by Goldberger and Olkin's analysis of *LIML*. Therefore minimizing Q is equivalent to minimizing the *GLS* criterion $G = tr(S^{-1}W)$.

The parameters π can be obtained directly from the likelihood function, as described by Jöreskog, by iteration. The standard errors of the parameters are then obtained from the second partial derivatives of G, evaluated at the minimum value. However, π can also be found, without iteration, by differentiating G with respect to each parameter in turn.

The elements of S^{-1} are S^{CC}, S^{CY}, and S^{YY}; except for the constant $det(S^{-1}) = (det\, S)^{-1}$, these are equal respectively to S_{YY}, $-S_{CY}$, and S_{CC}, which are the *OLS* estimators of V_{YY}, $-V_{CY}$ and V_{CC} in (11). Then the estimating criterion G can be written

$$S_{YY}(C - H\theta - \mu X\phi)'(C - H\theta - \mu X\phi) - 2S_{CY}(C - H\theta - \mu X\phi)'(Y - X\phi) + S_{CC}(Y - X\phi)'(Y - X\phi)$$

In order to separate the direct effects (θ) from the indirect effects ($\mu\phi_H$) of the H variables, any correlation between H and Y, C and Z must be removed. This is accomplished by the transformations

$$\tilde{Y} = [I - H(H'H)^{-1}H']Y = \Lambda Y = Y - H\psi_Y$$
$$\tilde{C} = [I - H(H'H)^{-1}H']C = \Lambda C = C - H\psi_C$$
$$\tilde{Z} = [I - H(H'H)^{-1}H']Z = \Lambda Z = Z - H\psi_Z$$

where ψ_Y and ψ_C are $r \times 1$, and ψ_Z is $r \times (m - r)$. Since $\tilde{H} = 0$, $\tilde{X} = [0, \tilde{Z}]$. Therefore $r < m$ is required for identification, or not all variables in X can be allowed to affect the propensity to consume.

Successive differentiation of G with respect to each of the parameters and substitution into the criterion leads to a quartic in μ. Given estimates of μ and of S, the remaining parameters can be expressed as functions of observed variables and the transformations Λ or ψ. Thus $\phi_Z = (Z'\Lambda Z)^{-1}Z'\Lambda[\alpha_y Y + \alpha_c C]$ where the weights α_y, α_c applied to Y and C are

$$\alpha_Y = (\mu^2 S_{YY} - 2\mu S_{CY} - S_{CC})^{-1}(S_{CC} - \mu S_{CY})$$
$$\alpha_C = (\mu^2 S_{YY} - 2\mu S_{CY} - S_{CC})^{-1}(\mu S_{YY} - S_{CY})$$

with their sum $\gtrless 1$ as $\mu \lessgtr 1$. The relative importance of observed income and consumption in determining permanent income thus turns on the relative sizes of α_Y and α_C. Note that the conditions $\alpha_Y \geq 0$, $\alpha_C \geq 0$, correspond to the limits S_{CC}/S_{CY} and S_{CY}/S_{YY} for the parameter μ. The vector ϕ_Z can also be written, less compactly, in terms of the transformation ψ_Z:

$$\phi_Z = (Z'(Z - H\psi_Z)Z)^{-1}Z'(Z - H\psi_Z) \cdot [\alpha_Y Y + \alpha_C C]$$

The use of ψ is more convenient for the vectors ϕ_H and θ: the income effect of the H variables is $\phi_H = \psi_Y - \psi_Z\phi_Z = (H'H)^{-1}H'[Y - Z\phi_Z]$ while their effect on the propensity to consume is $\theta = \psi_C - \mu\psi_Y = (H'H)^{-1}H'(C - \mu Y)$. Note that ϕ_H depends on ϕ_Z ($Y - Z\phi_Z$ is an estimate of the part of income not accounted for by Z), while θ depends only on μ.

The regressions for ϕ and θ provide estimates of the standard errors of the parameters, but these are biased toward zero because they are estimated sequentially rather than simultaneously. No account is taken of the errors in estimating μ, for example, when ϕ and θ are estimated. Exact standard errors can be estimated, however, by an F-test. Let G_V be the value of G obtained by minimizing with no restriction on the parameters, and let G_V^R be the value obtained under linear restrictions on one or more parameters. If f parameters are restricted, the statistic

$$\frac{(T-m-r-1)}{f} \cdot \frac{(G_V^R - G_V)}{G_V}$$

is distributed as F, with f and $T-m-r-1$ degrees of freedom, T being the number of observations. When $f=1$, a single parameter ξ is restricted to the value ξ^R. The true standard error is then given by $\sigma(\xi) = |\xi^R - \xi|/\sqrt{F}$.

Two kinds of restriction are of particular interest. One is that $\mu = 1$, to test the proportionality hypothesis in the logarithmic version. The other restriction is that one or more variables have no effect on permanent income or on the propensity to consume. In the linear version of the PIH, the proportionality assumption is $H\theta = 0$. If variables in Z are restricted ($\phi_{Z_i} = 0$ for some i), H is unchanged, so $\tilde{Z}$ is simply a subset of the unrestricted $\tilde{Z}$, with Λ or ψ_Z fixed. If variables are omitted from H, the procedure is complicated by the necessity to reestimate Λ. The hypothesis that $\theta_i = 0 \neq \phi_{H_i}$, shifts the variable H_i from H to Z. The reverse assumption $\phi_{H_i} = 0 \neq \theta_i$ removes H_i from X altogether.

REFERENCES

J. Aitchison and J. A. C. Brown, *The Lognormal Distribution*, Cambridge 1963.

C. L. F. Attfield, "Estimation of the Structural Parameters in a Permanent Income Model," *Economica*, Aug. 1976, *43*, 247–54.

R. A. Berry, "Open Unemployment as a Social Problem in Urban Colombia: Myth and Reality," *Econ. Develop. Cult. Change*, Jan. 1975, *23*, 276–91.

R. R. Betancourt, "The Normal Income Hypothesis in Chile," *J. Amer. Statist. Assn.*, June 1971, *66*, 258–63.

———, "The Analysis of Patterns of Consumption in Underdeveloped Countries," paper presented to the ECIEL Conference on Consumption, Prices and Economic Development, Hamburg, Germany, Oct. 1973.

S. S. Bhalla, "The Measurement of Permanent Income and its Application to Savings Behavior," unpublished paper, World Bank 1978.

———, "Measurement Errors and the Permanent Income Hypothesis: Evidence from Rural India," *Amer. Econ. Rev.*, June 1979, *69*, 295–307.

J. Crockett and I. Friend, "Consumption and Saving in Economic Development," work. paper no. 22–73, Rodney L. White Center Finan. Res., Univ. Pennsylvania.

R. Ferber, "Consumer Economics, A Survey," *J. Econ. Lit.*, Dec. 1973, *11*, 1303–42.

Milton Friedman, *A Theory of the Consumption Function*, Princeton 1957.

A. S. Goldberger, "Maximun-Likelihood Estimation of Regressions Containing Unobservable Independent Variables," *Inter. Econ. Rev.*, Feb. 1972, *13*, 1–15.

——— **and I. Olkin,** "A Minimum-Distance Interpretation of Limited-Information Estimation," *Econometrica*, May 1971, *39*, 635–40.

H. J. Howe and P. Musgrove, "An Analysis of ECIEL Household Budget Data for Bogotá, Caracas, Guayaquil and Lima," in Constantino Lluch, Alan Powell and Ross Williams, eds., *Patterns in Household Demand and Saving*, New York; Oxford 1977.

K. G. Jöreskog, "A General Method for Estimating a Linear Structural Equation System," in Authur S. Goldberger and Otis D. Duncan, eds., *Structural Equation Models in the Social Sciences*, New York 1973.

A. C. Kelley and J. G. Williamson, "Household Saving Behavior in the Developing Economies: The Indonesian Case," *Econ. Devel-*

op. Cult. Change, Apr. 1968, *16*, 385–403.

S. Kuznets, "Demographic Aspects of the Size Distribution of Income: An Exploratory Essay," *Econ. Develop. Cult. Change*, Oct. 1976, *25*, 1–94.

M. Lansberger, "Windfall Income and Consumption: Comment," *Amer. Econ. Rev.*, June 1966, *56*, 534–40.

Thomas Mayer, *Permanent Income, Wealth and Consumption*, Los Angeles 1972.

F. Modigliani and A. Ando, "The Permanent Income and the 'Life Cycle' Hypothesis of Saving Behavior: Comparison and Tests," in Irwin Friend and Ronald Jones, eds., *Consumption and Saving*, Philadelphia 1960.

Philip Musgrove, "Determination and Distribution of Permanent Household Income in Urban South America," unpublished doctoral dissertation, M.I.T., 1974.

———, "An Extended Linear Permanent Expenditure System," in Alan S. Blinder and Philip Friedman, eds., *Natural Resources, Uncertainty, Dynamics and Trade: Essays in Memory of Rafael Lusky*, New York 1977.

———, (1978a) "Determinants of Urban Household Consumption in Latin America: A Summary of Evidence from the ECIEL Surveys," *Econ. Develop. Cult. Change*, Apr. 1978, *26*, 441–65.

———, (1978b) *Consumer Behavior in Latin America*, Washington 1978.

A. M. Okun, "The Personal Tax Surcharge and Consumer Demand, 1968–70," *Brookings Papers*, Washington 1971, *1*, 167–204.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

R. Ramanathan, "Estimating the Permanent Income of a Household: An Experiment in Methodology," *J. Polit. Econ.*, Jan./Feb. 1971, *79*, 177–85.

V. Stoikov, "How Misleading Are Income Distributions?," *Rev. Income Wealth*, June 1975, *21*, 239–50.

F. Vakil, "The Propensity to Consume Permanent Income in India," *Econ. Develop Cult. Change*, Apr: 1973, *21*, 514–21.

A. Zellner, "Estimation of Regression Relationships Containing Unobservable Independent Variables," *Int. Econ. Rev.*, Oct. 1970, *11*, 441–54.

Welfare, Remarriage, and Marital Search

By ROBERT M. HUTCHENS*

Between 1964 and 1974 female-headed families with children under 18 increased as a proportion of all *U.S.* families by 40 percent. The Aid to Families with Dependent Children (*AFDC*) caseload more than doubled. Though increased female headship undoubtedly contributed to caseload growth, the *AFDC* program's structure may have simultaneously encouraged female headship. By focusing payments on families without husbands, the program could conceivably promote marital separations and discourage remarriage, thereby contributing to the growth of female-headed families.[1] The testing of such hypotheses bears on a number of policy issues. We need to understand the degree to which the *AFDC* program's goal of alleviating the financial problems of female-headed families induces the formation and continuation of such families. In addition, we need to determine the extent to which growth in the *AFDC* caseload is caused by these separation and remarriage incentives. Finally, we need to know whether alternative support programs could reduce the problem.

This paper focuses upon one aspect of this complex issue: the effect of *AFDC* transfers on the probability of remarriage for female heads with children. In the first section search theory is applied to marriage, establishing hypotheses on the links between remarriage, *AFDC* transfers, and other observable socioeconomic variables. The second section presents empirical tests of these hypotheses.

I. Theoretical Determinants of Remarriage

Remarriage is a voluntary act undertaken in a world of imperfect information, requiring an investment of time and money in search of potential marriage partners. As such, remarriage can be analyzed through a theory of search. The theoretical literature on search focuses in large part on the labor market, though a parallel application to marital behavior is beginning to appear.[2] The parallels are striking.

In its simplest form the modern theory of labor market search assumes the individual confronts a known distribution of wage offers arising from differing employer valuations of the individual's skills. To obtain information

*Assistant professor, New York State School of Industrial and Labor Relations, Cornell University. I have benefited from the suggestions of Harold Watts, Glen Cain, Irwin Garfinkel, Ronald Ehrenberg, Larry Orr, Jon Christianson, Maurice MacDonald, the managing editor, and an anonymous referee. Financial support was provided by the Institute for Research on Poverty, University of Wisconsin, and the New York State School of Industrial and Labor Relations.

[1]The question has been addressed by Marjorie Honig, Heather Ross and Isabel Sawhill; Sawhill et al.; Greg Duncan. Honig utilized 1960 and 1970 data on forty-four Standard Metropolitan Statistical Areas (*SMSAs*) to examine the relationship between *AFDC* payments and the proportion of women in the *SMSA* who are female heads with children. The estimated relationship was generally positive. Her work does not permit analysis of the proportion of welfare-related female heads attributable to marital disruption, lower rates of remarriage, migration of female heads to the high benefit *SMSAs*, or formation of own households by female heads. Using 1970 Census data on low-income areas in forty-one *SMSAs*, Ross and Sawhill confirmed Honig's results for blacks but not for whites. Sawhill et al. utilized data from the Office of Economic Opportunity (*OEO*) *Panel Study of Income Dynamics* to analyze the relationship between *AFDC* payments and separation and remarriage probabilities over four years. They found no significant relationship between *AFDC* payment levels and such transitions, but did establish that *AFDC* receipients are less likely to remarry than other female heads, *ceteris paribus*. Duncan also worked with the *OEO* panel in analyzing remarriage over six years. He found virtually no evidence of a consistent relationship between *AFDC* payments and remarriage. Finally, the negative income tax experiments have provided additional insights into the relationship between cash transfers and marital transition. For a review of this literature and some interesting new research, see Michael Hannan, Nancy Tuma, and Lyle Groeneveld.

[2]See Gary Becker, Elizabeth Landes, and Robert Michael. Their work focuses on marital search as part of a lifetime utility-maximization strategy, and encompasses marriage, separation, remarriage, and fecundity. As such it does not explicitly deal with the issues addressed here.

on wage offers, the individual randomly samples the offer distribution and thereby incurs costs such as the opportunity cost of search time. Assuming risk neutrality, he (she) seeks to maximize his (her) expected discounted lifetime earnings (net of search costs) by deciding whether to participate in search, and upon undertaking search deciding when to accept an offer. The more selective his (her) criteria for an acceptable offer, the longer will be his (her) expected duration of search. The theoretical literature on search considers the behavioral implications of alternative assumptions regarding such factors as time horizons, rights of recall, and maximization of utility vs. income. Essential hypotheses regarding participation in and duration of search appear to be invariant to many of the assumptions. Specifically, holding other things equal:[3]

1) Variables which increase the cost of search tend to reduce the probability of participation in search and reduce the duration of search.

2) Variables which increase the expected flow of "nonportable" income going to searchers and nonparticipants reduce the probability of participation and increase duration. Nonportable income terminates upon acceptance of an offer. Unemployment insurance is an example of such income.

3) Variables which increase the individual's discount rate tend to decrease the probability of participation and decrease duration.

4) Variables which increase the individual's time horizon tend to increase the probability of participation and increase duration.

5) Variables which increase the mean of the offer distribution tend to increase the probability of participation with ambiguous effects on duration.

Marital search can be considered through a parallel framework. Once again, assume that the individual faces a known distribution of offers and incurs costs in obtaining information on specific offers.[4] The offer can be viewed as an expected stream of real income or commodities, with commodities assumed to be produced by households, to enter utility functions, and to include "... the quality of meals, the quality and quantity of children, prestige, recreation, companionship, love, and health status" (Becker, 1973, p. 816). Differences in offers are caused by differences in the value which potential marriage partners give to the individual's characteristics.[5] Again, assuming risk neutrality, the individual seeks to maximize the net benefits from search by deciding whether to undertake search and whether to stop searching and accept an offer. When marriage is conceptualized in this way, the hypotheses listed above are directly applicable to marital search.

A complete specification of a marital search model should include measures of the individual's cost of search, expected future flow of real income when single, discount rate, time horizon, and the distribution of offers. Empirical research confronts the considerable difficulty of operationalizing such theoretical constructs. As with most research in this area, I shall not attempt to test a complete model of search. Rather, the empirical work is restricted to testing hypotheses for observed characteristics of potential searchers, with the

[3]See Steven Lippman and John McCall; Kenneth Burdett; Ronald Ehrenberg and Ronald Oaxaca for reviews of the literature. See also Reuben Gronau and Dale Mortensen for discussions of participation in search.

[4]Examples of activities associated with gathering such information are dating, visits to a singles bar, and attendance at dances. These activities could have benefits in and of themselves which are not associated with search. To avoid complicating the analysis, it is implicitly assumed that the costs of obtaining information exceed the pleasures arising from such activities in and of themselves.

[5]More precisely, following Becker let the household produce one abstract commodity Z, with the level of output in a marriage between the ith female and the jth male denoted by Z_{ij}. Assume in addition that Z_{ij} is divided between i and j such that $Z_{ij} = Z^m_{ij} + Z^f_{ij}$, where the superscripts denote male and female shares of the household output. The jth male will then be willing to make an offer of Z^f_{ij} to the ith female which satisfies $Z^f_{ij} \leq Z_{ij} - Z_j^{m*}$, where Z_j^{m*} is the highest level of real income attainable in other marriages for the jth male. The male's offer then makes the male at least as well off as he would be in other marriages. Thus, the jth male's offer depends on Z_{ij} (which in turn depends on the unique characteristics of male j and female i) and Z_j^{m*}. Since Z_{ij} and Z_j^{m*} will vary across potential marriage partners, the level of offers will also vary.

distribution of offers and the individual rate of discount remaining largely unobserved.

The first hypothesis relates an increase in *AFDC* transfers to one female head's remarriage behavior. In most cases such transfers terminate with marriage,[6] and consequently do not affect the distribution of marriage offers.[7] To the extent that the female head perceives a finite probability of receiving such income, the increase tends to raise the expected future flow of income when single, thereby reducing the probability of participation in search and increasing the duration of search. Alternatively, given imperfect capital markets, *AFDC* transfers could reduce the individual's discount rate, and thus increase both the probability of participation in and the duration of search. An increase in *AFDC* payments should then unambiguously increase the duration of search, and, assuming a minor impact on the time rate of discount, decrease the probability of participation in search.

The relationship between marital search behavior and increased portable nonwage income (income which does not terminate upon marriage such as property income) is considerably more complicated. As with *AFDC* transfers the increase may 1) reduce the individual's discount rate and 2) raise the expected flow of income when single. In addition, it enhances the woman's attractiveness to potential spouses, causing 3) an increase in the mean of the offer distribution. Effects 1 and 2 increase the duration of search and effect 3 has an indeterminate impact on duration. Effects 1 and 3 increase the probability of participation in search and effect 2 reduces the probability. Becker's theoretical analysis indicates that the third effect tends to dominate the second effect, that is, the rise in real income implicit in the mean marriage offer exceeds the rise in real income implicit in remaining single, causing an increase in the gain from marriage (see Becker, 1973, p. 821). Thus there are solid reasons to expect greater portable nonwage income to increase both the duration of search and the probability of participation in search.

A rise in the female head's wage rate similarly affects marital search in several ways. As with portable nonwage income, the change 1) raises the flow of income when single, 2) may reduce the individual's discount rate, and 3) increases attractiveness to potential spouses causing a rise in the mean of the offer distribution. In addition, increased wages could 4) raise the opportunity cost of time allocated to search. Higher wages then have an ambiguous impact on duration of search since effects 1 and 2 increase duration, effect 4 reduces duration, and the impact of 3 is indeterminate. The relationship between higher wages and participation probabilities is similarly ambiguous since effects 2 and 3 increase the probability, and effects 1 and 4 reduce the probability.[8] Higher wages thus have an indeterminate impact on marital search behavior.

Increased numbers of children and presence of young children may increase the opportunity cost of time spent in search, and reduce the mean of the offer distribution by lowering the woman's attractiveness as a potential partner. Since the first effect unambiguously decreases duration while the impact of the second effect is indeterminate, one may reasonably hypothesize that such variables reduce duration of search. The probability of participation in search should also fall since both effects reduce this probability.

[6]If the woman marries a man eligible for *AFDC*-Unemployed Father or a disabled male, she could continue receiving *AFDC* payments. In addition, a 1970 Supreme Court decision (Lewis vs. Martin) declared that in absence of proof of actual contribution, states cannot consider the resources of a nonadopting stepfather or other male in the home in determining payments. Given normal delays in implementation, it can be assumed that this did not affect the empirical work presented here.

[7]When additional women in the marriage market receive a similar increase in *AFDC* transfers, the offer distribution could change. *AFDC*-induced withdrawals from the marriage market could drive up the level of the average offer confronting a given woman.

[8]Becker's theory (1973, p 822) would suggest that if the wage of all potential partners were below (above) the female head's wage, an increase in her wage would tend to increase (decrease) the gain from marriage. Under these very restrictive assumptions one could argue that the third effect will tend to dominate (be dominated by) the first effect.

Assuming that time horizons become shorter with age, increased age of the female head reduces both the probability of undertaking search and the duration of search. Becker's work can be interpreted as reinforcing such conclusions. The main advantage of marriage over being single, according to Becker, "... lies in the desire to raise own children and the physical and emotional attraction between sexes. Nothing distinguishes married households more from singles households or from those with several members of the same sex than the presence, even indirectly, of children" (1973, p. 818). From this Becker is able to claim that the gain from marriage is positively related to the importance of own children. Since aging is associated with already having produced own children, with decline in reproductive capabilities, and with concern over the possibility of dying while own children are young, we would expect the importance of own children to decline with increased age. Such considerations imply that the real income implicit in marriage offers declines with increases in the woman's age.

The ratio of males to females in the relevant marriage market may also affect marital search behavior. A rise in this ratio increases the mean of the offer distribution[9] and therefore increases the probability of participation in search. The impact on duration of search is, however, indeterminate.

Clearly, other variables such as education, religion, race, health, residence, and physical appearance play a role in marital search behavior. Theoretical effects depend upon characteristics of potential marriage partners and the extent to which heterogeneity of traits is desired in marriage. A full analysis of these issues is beyond the scope of this paper.

II. Empirical Testing of Hypotheses

The empirical work utilizes data on female heads with children drawn from the "Michigan" data (*OEO Panel Study of Income Dynamics*), and is restricted to twenty states with large *AFDC* populations.[10] Hypotheses were tested through maximum likelihood estimation of a logistic model of the form[11]

$$P = 1/(1 + e^{-BX})$$

where P is the probability of remarriage over two years, X is a vector of exogenous variables, and B is a vector of estimated coefficients.

An ideal test of all search-related hypotheses would utilize information on participation in and duration of marital search. Unfortunately a data set containing such information is not presently available.[12] In general we must investigate determinants of the probability of marital transitions and recognize that people who do not experience transitions may be either long-term searchers or nonparticipants in search.

Since the primary purpose of this paper is to analyze the relationship between *AFDC* transfers and remarriage, probabilities are measured over a relatively short period of time (two years). It was hypothesized that increased *AFDC* transfers reduce participation in search and lengthen the expected duration of search by increasing selectivity in

[9] As noted in fn. 5, the jth male's marriage offer to the ith female must satisfy $Z^f_{ij} \leq Z_{ij} - Z^{m*}_j$. An increase in the ratio of males to females will tend to reduce Z^{m*}_j (see Becker 1973, p. 840; Alan Freiden, pp. S34–37), which leads in turn to a rise in the upper bound on the jth male's offer to the ith female.

[10] The twenty states were: Alabama, California, Florida, Georgia, Illinois, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Mississippi, Missouri, New Jersey, New York, North Carolina, Ohio, Pennsylvania, Tennessee, Texas, and Washington. In 1970, 79 percent of the *AFDC* recipient population lived within these states.

[11] For a discussion of maximum likelihood estimation of the logistic function see Marc Nerlove and S. James Press.

[12] Such a data set is, however, conceivable. Assume a body of data on individual lifetimes and containing measures of all independent variables. Participants in search would be defined as people who, having separated, remarried at some time in their lives. Participation hypotheses would then be tested in the population which separated, and duration of search hypotheses would be tested in the population which participated. Admittedly a measurement problem arises from individuals participating in search yet never marrying. Participation hypotheses are then tested in a sample containing an error in the dependent variable, and duration hypotheses are tested in a truncated sample. Both problems are potentially resolvable given appropriate econometric techniques. A meaningful empirical test of a theory which distinguishes marital participation from duration of search is then possible.

TABLE 1—HYPOTHESIZED RELATIONSHIPS BETWEEN EXOGENOUS VARIABLES AND SEARCH PARTICIPATION, SEARCH DURATION, AND TWO-YEAR REMARRIAGE PROBABILITIES

Variable	Probability of Partici-pation in Search	Duration of Search	Expected Impact on Remarriage Probability over Two Years
AFDC Transfers	-	+	-
Portable Nonwage Income	+	+	?
Wage Rate	?	?	?
Number of Children and Presence of Young Children	-	-	?
Age		-	?
Sex Ratio	+	?	+

choosing a marriage partner. Though both effects reduce the probability of an observed remarriage during a given time interval, shortening the interval over which the probabilities are measured tends to amplify the duration effect of *AFDC* transfers.[11]

While short measurement periods facilitate analysis of *AFDC* transfers, they confuse the analysis of other variables. Table 1 summarizes hypotheses presented in the previous section and notes each variable's expected impact on remarriage probabilities over two years. Portable nonwage income illustrates the problems inherent in testing hypotheses with such data. Theoretically, such nonwage income tends to both increase participation in and duration of search. Over a short period the participation effect will tend to increase remarriage probabilities while the duration effect (which is due to increased selectivity in choosing a marriage partner) will tend to reduce remarriage probabilities. Since we have no a priori expectations on which effect will dominate, hypotheses are indeterminate. As indicated in Table 1 the only explicit hypotheses for the short-period analysis concern *AFDC* transfers and sex ratios. Even the sex ratio hypothesis is somewhat ambiguous since higher sex ratios could conceivably lead to greater duration of search.

A. *Variables*

The following variables were included in the analysis of remarriage probabilities:

Remarriage: A woman with children, who is classified as a female head in year t, is considered remarried if she lives in a male-headed household and can reasonably be classified as the wife of the head in year t + 1 or t + 2.[14] This classification is consistent with the Becker model where two people of the opposite sex sharing the same household are considered married. Note that reconciliation is then treated as a form of remarriage.

AFDC transfers: Two measures are used to describe the monthly amount of funds available from the *AFDC* system: the guarantee and the break-even income level.[15] Given a woman's nonwage income, the guarantee is the level of payments going to a woman with zero earned income. The break-even income level is the amount of earned income at which *AFDC* payments fall to zero. The variables are constructed from published data on the *AFDC* system, the 1967 and 1971 *AFDC* surveys, and *OEO* panel data on family size, location, and nonwage income.[16] In addition, a binary variable indicating receipt of *AFDC*

[11]An empirical test on probabilities over one year yielded *AFDC* transfer coefficients which were quite similar to the two-year coefficients Given the small number of observations on the dependent variable this would appear to be a very weak test of the duration effect

[14]To determine whether the female head in period t can reasonably be classified as the wife of the head, the age of the wife in t + 1 or t + 2 is compared to the age of the female head in t

[15]This paper focuses on *AFDC* program parameters and does not therefore deal with in-kind programs such as vocational training, public housing, food stamps, and Medicaid. Extension of remarriage analysis to such programs will require measures of the cash equivalent value of such benefits at the individual level. During the period under consideration, *AFDC* benefits were the primary form of income-tested transfers available to female-headed families.

[16]Given the kinked form of the *AFDC* income-leisure constraint, it should be described with three parameters. The appropriate parameterization of the constraint is discussed in my 1976 paper. Since two of the three parameters are highly correlated in this sample (.96), a two-parameter model is used here. The parameters are computed as (over)

payments during year t is utilized. Finally, since *AFDC* transfers should only influence marital transitions when there is a finite perceived probability of receipt, the guarantee and breakeven were interacted with a predicted probability of *AFDC* receipt in 1970.[17]

Nonwage income: The *OEO* panel provides statistics on annual nonwage income. In general nonwage income is measured in t + 1 for female heads at risk of remarriage in t + 1 or t + 2.[18] The measure used here includes income from rent, interest, dividends, and transfers, plus an imputed asset portion of income from farm, business, roomers, and boarders. The transfer component is important because it may contain nonportable elements. Included in transfers are Social Security payments, income from retirement pay, pensions and annuities, unemployment compensation, worker's compensation, alimony, child support, income from relatives, and income from other sources (besides earnings).[19] If the individual receives alimony or Social Security survivor's benefits, her level of nonwage income will unambiguously fall when she legally remarries.[20] In addition, worker's compensation death benefits, survivor's benefits from employee benefit plans, and transfers from relatives may fall upon remarriage. Thus, the nonwage income measured here may not be completely portable between single and married status. Since it was not possible to separate portable from nonportable elements, nonwage income was disaggregated and the impact of its several components analyzed.

Hourly wage: A five-year average of wages is used as a measure of the gross wage offered a female head.[21] As opposed to a

$$\text{Guarantee} = M \text{ if } M < r(pS - NWY)$$
$$= r(pS - NWY) \text{ if } M > r(pS - NWY)$$
$$\text{Breakeven} = r(pS - NWY + K)/AFDC \text{ tax rate}$$

where S is the state's standard of need for a family of given size; p is a percentage by which the standard of need (S) is reduced for purposes of calculating benefits. For most states p has a value of 1, and is often termed a percentage reduction; NWY is unearned income for the family; M is a maximum payment level for a family of given size; r is a percentage applied to the difference between needs (pS) and family resources (income minus deductions). For most states r has a value of 1 and is often termed a ratable reduction; K equals zero before July 1969, and $20 thereafter. The guarantee is constrained to be greater than or equal to zero. NWY is obtained from the *OEO* panel data. Data on p, M, and r are obtained from Department of Health, Education, and Welfare publications. *AFDC* tax rates are presented in my 1978 paper. Since the *OEO* panel data provide no information on S, this variable had to be obtained from external sources. To accomplish this, regressions were estimated with 1967 and 1971 *AFDC* survey data in each of the twenty states using female *AFDC* recipients as the unit of observation. The dependent variable was the standard of need employed by the welfare agency in computing the family's payments. Independent variables included family size, published data on the state needs standard for two- and four-member families, and a binary variable indicating urban residence. The estimated regressions were then used to predict the level of S for each female head in the twenty-state Michigan sample. For a more detailed discussion see my 1976 paper, Appendix D.

[17]The probability of *AFDC* receipt in 1970 was estimated with a logistic function. The estimated model is:

$$\log_e P/(1-P) = \underset{(1.2213)}{.77360} - \underset{(2.5355)}{.35616X_1} - \underset{(2.0806)}{.00300X_2} + \underset{(2.8172)}{.005143X_3} + \underset{(.25281)}{.00008X_4} - \underset{(.51559)}{.20154X_5} - \underset{(2.776)}{.03455X_6} + \underset{(1.2679)}{.03715X_7} + \underset{(3.3910)}{.8967X_8}$$

where P is the probability of *AFDC* receipt, X_1 is the wage rate, X_2 is nonwage income, X_3 is the *AFDC* guarantee, X_4 is the *AFDC* breakeven, X_5 is no earnings binary variable, X_6 is the age of female head, X_7 is a binary variable indicating presence of children under five, X_8 is a binary variable indicating disability, and t-statistics are in parentheses. In general the variables included in the model were statistically significant in an analysis of *AFDC* entry and exit reported in my 1976 paper.

[18]Measurement of nonwage income is complicated by a timing problem. Nonwage income during year t + 1 is reported in the next year's interview. If the woman married during t + 1, the Michigan data measures unearned income of both the husband and wife in the next year's interview. In this case, the female head's nonwage income in year t (the t + 1 interview) was used.

[19]Specifically excluded are *AFDC* transfers and other welfare payments. These are quite likely to be nonportable.

[20]Remarriage by a widow under 60 leads to loss of *her* survivorship benefits, though the children's benefits continue.

[21]Specifically, the wage used in year t was

$$\sum_{i=1}^{5} \frac{(Earnings(i)/(CPI(i) \times Hours(i)))}{N} \times CPI(t)$$

predicted wage, this average should reflect all human capital, geographic, and individual determinants of the wage confronting a given woman. For women with no earnings over the five study years, the wage is set at zero and a binary variable (named No Earnings) set equal to one.

Age of youngest child: A binary variable takes the value 1 when the youngest child is under 5.

Number of children: This is the number of children under 18 living with the mother.

Sex ratio (the ratio of males to females). 1970 census data on the ratio of males to females in each female-head's county and age classifications (20–25, 26–30, etc.) were used for this variable. More precise data on sex ratios by age, county, race, and marital status were not available.

Additional variables were used as controls. In general the selected variables appeared to be important in previous works on remarriage. Included are race, disability, marital status, location (region and urban/rural), and female headship in the previous year.

B. *Results*

Table 2 presents parameter estimates from logistic models estimated on a population of 1970 female heads. All results are consistent with the hypothesis that *AFDC* transfers reduce the probability of remarriage.[22] The first model parameterizes the *AFDC* "offer" with the guarantee level. The resultant negative coefficient is statistically significant (one-tail test) at a .025 level. To allay any doubt that *AFDC* transfers are in fact influencing remarriage, consider models (2) and (3). In model (2) the *AFDC* guarantee is replaced by a binary variable indicating receipt of *AFDC* transfers during 1970. This is negative and statistically significant. Model (3) interacts the *AFDC* guarantee with an estimated probability of receiving *AFDC* transfers (for details see fn. 19). This specification attempts to operationalize the hypothesis that changes in the guarantee should only affect behavior when there is a finite probability of receiving *AFDC* payments. Note that the statistical significance of this parameter exceeds that of the simple guarantee in model (1).[23] The fourth model adds the breakeven to the parameterization of the *AFDC* offer. Though both the guarantee and breakeven are negative, neither is statistically significant due to the relatively high degree of correlation between the two variables (+.837).

The results can be illustrated with a simple example. For a family of four (a woman and three children) the 1971 guarantee in New York was $305 while the guarantee in Mississippi was $60. If a Mississippi woman with three children had a .50 probability of remarriage, what effect would Mississippi's adoption of the New York guarantee have on her probability of remarriage? From model (1) the $245 increase in the guarantee would alter the linear combination of variables and coefficients by −1.76. This translates into a postchange remarriage probability of .147.

---

where *Earnings*(i) is the woman's gross earnings in the ith year; *Hours*(i) is the woman's hours of labor in the ith year; *CPI*(i) is the Consumer Price Index in the ith year, N is the number of years in which the woman had nonzero earnings.

[22]Duncan and Sawhill et al. also used the Michigan data to examine the relationship between marriage and the level of *AFDC* payments, but failed to obtain meaningful results. Three factors may have contributed to the difference in results. 1) This work uses external sources to measure the *AFDC* guarantee and imputes a guarantee to all female heads in the sample. Both Duncan and Sawhill et al. employ the actual amount of *AFDC* payments received during 1967 as their measure of the *AFDC* offer. Their measure is affected by individual wage income and months of female headship in 1967. In addition, their specification implicitly assumes that the only female heads influenced by the *AFDC* system are those who actually received *AFDC* in 1967. 2) In this work the probability of remarriage is measured over two years. Duncan measures the probability over six years and Sawhill measures it over four years. As noted in the text, there are theoretical grounds for expecting shorter time periods to amplify the impact of the *AFDC* guarantee. 3) This analysis is restricted to women who were female heads in 1970, whereas Duncan and Sawhill et al. consider women who were female heads in 1968. As noted in the text, the Michigan data suffer serious problems with sample attrition between the 1968 and 1970 interviews. By focusing on 1970 these problems are substantially reduced.

[23]It can be argued that a likelihood ratio test is more appropriate for testing hypotheses in a logistic model. Use of this test indicated that the coefficient on "Guarantee x Probability of Receipt" is significant at a .025 level.

TABLE 2—LOGISTIC ANALYSIS OF THE PROBABILITY OF REMARRIAGE OVER TWO YEARS FOR 1970 FEMALE HEADS WITH CHILDREN[a]

Variable	Model (1)	(2)	(3)	(4)	(5)
Guarantee	−0.0072	–	–	–	–
	(2.0)				
Receipt of *AFDC*	–	−1.3997	–	–	–
		(3.0)			
Guarantee x Probability of Receipt	–	–	−0.0096	−0.0072	−0.0108
			(2.2)	(1.1)	(2.3)
Breakeven x Probability of Receipt	–	–	–	−0.0005	–
				(.5)	
Age	−0.0955	−0.1018	−0.1053	−0.1068	−0.1031
	(3.4)	(3.5)	(3.6)	(3.6)	(3.5)
Sex Ratio	0.4625	0.1992	0.3829	0.3912	0.4616
	(0.3)	(0.1)	(0.3)	(0.3)	(0.3)
Nonwage Income	−0.0042	−0.0019	−0.0035	−0.0036	–
	(1.7)	(1.2)	(1.7)	(1.7)	
Hourly Wage	−0.2355	−0.3275	−0.3478	−0.3506	−0.3079
	(1.0)	(1.3)	(1.4)	(1.4)	(1.2)
No Earnings	−0.4355	−0.6463	−0.4413	−0.4740	−0.3721
	(0.6)	(0.9)	(0.6)	(0.6)	(0.5)
Number of Children	0.1338	0.0228	0.1298	0.1321	0.1550
	(0.9)	(0.2)	(1.0)	(1.0)	(1.1)
Presence of Young Children	−0.3487	−0.2187	−0.4099	−0.4200	−0.3616
	(0.8)	(0.6)	(0.9)	(1.0)	(0.8)
Nonwhite	−1.455	−1.5374	−1.4313	−1.4093	−1.5402
	(3.2)	(3.3)	(3.2)	(3.1)	(3.4)
Disability	0.0354	0.1689	0.2576	0.2642	0.2138
	(0.1)	(0.4)	(0.6)	(0.6)	(0.5)
Separated	0.2259	0.2430	0.2359	0.2354	0.1476
	(0.5)	(0.6)	(0.6)	(0.6)	(0.3)
Widowed	−0.2236	−0.2893	−0.1353	−0.0964	0.0258
	(0.3)	(0.4)	(0.2)	(0.1)	(0.0)
Female Head in t−1	−1.0648	−0.9130	−1.1029	−1.1115	−1.0179
	(2.0)	(1.7)	(2.1)	(2.1)	(1.8)
South	0.1034	0.3056	0.0936	0.1107	0.0164
	(0.2)	(0.6)	(0.2)	(0.2)	(0.0)
West	0.7824	0.5014	0.8310	1.0132	0.9589
	(1.4)	(0.9)	(1.5)	(1.5)	(1.7)
East	−1.2933	−2.0112	−1.3282	−1.3362	−1.3911
	(1.4)	(2.3)	(1.5)	(1.5)	(1.5)
Rural	0.1496	0.1342	0.1544	0.1547	0.0578
	(0.3)	(0.2)	(0.3)	(0.3)	(0.1)
Social Security Income	–	–	–	–	−0.0055
					(1.6)
Pension Income	–	–	–	–	−0.0042
					(1.0)
Unemployment and Worker's Compensation	–	–	–	–	−0.0542
					(1.0)
Alimony and Child Support	–	–	–	–	−0.0071
					(1.5)
Asset Income	–	–	–	–	−0.0038
					(0.6)
Other Income	–	–	–	–	−0.0007
					(0.3)
Constant	4.079	4.2363	4.3673	4.3656	4.3012
	(1.9)	(2.0)	(2.0)	(1.0)	(1.9)
Chi-Square	63.2	69.2	64.5	64.7	69.5
N	438	438	438	438	438

Source: Twenty-state subsample of the *OEO Panel Study of Income Dynamics*.

[a]Absolute t-statistics shown in parentheses.

Thus, for this particular woman the new guarantee would alter her probability of remarriage from .50 to .147.

From the theoretical discussion we expect increases in the ratio of males to females in the relevant marriage market to increase the probability of remarriage for female heads. The sign on the sex ratio coefficient is consistent with this hypothesis, but the coefficient is statistically insignificant. It is conceivable that a more precise measure of the sex ratio based on race, marital status, and education as well as age and location would support the theoretical hypothesis.

The theoretical analysis developed indeterminate hypotheses on the wage rate, number of children, presence of young children, and age of female head. A null hypothesis of a zero coefficient cannot be reasonably rejected for the first three variables, but can be rejected for age of female head. The result is consistent with the claim that aging has a major impact on participation in marital search.

Though theoretical hypotheses on nonwage income were indeterminate, the coefficient is negative and statistically significant at a .10 level (two-tailed test). As noted above, some components of the nonwage income variable may not be portable between single and married status, thus yielding an imperfect measure of the desired variable. To investigate the importance of this, nonwage income was disaggregated and its components introduced into the analysis. Model (5) in Table 2 presents the results. The coefficients on Social Security Income and Alimony and Child Support attain the highest level of statistical significance. To the extent that Social Security survivor's benefits and alimony are important sources of variation in the variables, the nonportable nature of these transfers could be generating the negative relationships. In consequence, these data do not permit conclusions on the relationship between portable nonwage income and remarriage. They are, however, consistent with findings on *AFDC* transfers, that is, nonportable transfers tend to reduce two-year remarriage probabilities.

The analysis also controlled for race, disability, location, marital status, and female headships in the previous year. A binary variable indicating race (1 when the head is nonwhite) was negative and highly significant; a result which is consistent with previous studies of remarriage. We may conclude that variation in remarriage probabilities by race does not arise from receipt of *AFDC* transfers.[24] The underlying reasons for this racial difference should be an important topic for future research. The pattern of coefficients on regional binary variables essentially follows that found in James Sweet's work, with the probability of remarriage being higher in the West and lower in the East than in the Central region.[25] The results on marital status were in general inconclusive.[26] Finally, a binary variable indicating female headship at the time of the 1969 interview (essentially acting as a control for duration of female-head status) was negative and statistically significant.

It should also be noted that several other control variables were tested in other logistic analyses. Included were education, local unemployment rate, savings, female head in $t-2$, and average wage of male unskilled workers in the female head's county. Given doubts on the accuracy of some of the measures as well as their failure to attain a meaningful level of significance, the variables were dropped from the model. They do not influence the conclusions presented here.

Two-year probabilities of remarriage can also be estimated for female heads in 1968 using models essentially identical to those in Table 2. When this was done, the coefficient estimates differed significantly from the 1970 estimates.[27] Though the 1970 estimates stand

[24]Further logistic analyses not presented here interacted the race variable with the *AFDC* guarantee. The resulting coefficient was statistically insignificant. Thus the null hypothesis that the two racial groups exhibit identical responses to the *AFDC* guarantee cannot be rejected.

[25]The South produces a minor anomaly. Sweet's work indicates that women in the South should have a lower probability of remarriage than those in the Central region.

[26]When both the "Widow" and "Separated" binary variables are zero, the female head is either single or divorced. The categories were combined since their legal status with respect to remarriage is essentially the same. In addition, four female heads with children who were reported as "married" were classified as "separated."

[27]Letting X_i denote the ith variable in Table 2 (thus X_3 is Guarantee x Probability of Receipt), the estimated

on their own, it is important to inquire into reasons for the difference. I would hypothesize that sample attrition seriously affects the 1968 and 1969 *Income Dynamics* remarriage data. In 1969 the Michigan interviewers were able to successfully reinterview only 89 percent of the 1968 families and 60 percent of the newly formed families. (See Institute for Social Research, pp. 26, 27.) By 1970 response rates had improved to 97 percent of the 1969 families and 84 percent of the newly formed families. Response rates again improved slightly in the 1971 and 1972 interviews. As noted by Isabel Sawhill et al.,

> ... attrition is probably biased so that more than a random distribution of female heads who have gotten (re)married are lost. This argument seems plausible since remarriage is probably associated with a higher probability of moving than no remarriage is and, with the name change which generally accompanies marriage, women may be more difficult to track down. [p. 139]

To the extent that attrition is differentially associated with specific socioeconomic variables, we may expect it to bias coefficients in the remarriage model. The problem would then be most serious in the 1968 sample when response rates were relatively low.[28]

Finally, from a policy perspective one would like to know the elasticity of remarriage with respect to changes in the guarantee. Accurate estimation of this elasticity requires data on the guarantee in all states, not simply the twenty states used here. Rather than undertake this considerable data collection task, I calculate an upper bound estimate of the elasticity based on the unweighted twenty-state sample of 1970 female heads. This is an upper bound estimate because the Michigan panel oversamples lower-income groups, i.e., it oversamples female heads likely to be influenced by the *AFDC* guarantee. Using model (1), the estimated elasticity is $-.8$. A 10 percent increase in the *AFDC* guarantee will then lead to no more than an 8 percent decline in remarriages by female heads with children (over 2 years), *ceteris paribus*. To illustrate, about 10 percent of the Michigan sample female heads remarried between 1970 and 1972 (unweighted). These results indicate that a 10 percent increase in the *AFDC* guarantee will change the remarriage rate from 10 percent to a number greater than 9.2 percent $(.10 - .08 \times .10 = .092)$.

III. Concluding Remarks

Theoretically, *AFDC* transfers should reduce the probability of participation in marital search and increase the duration of search for female heads with children. Both effects will tend to reduce the probability of remarriage over a short time period. The empirical work presented here indicates that an increase in the level of *AFDC* transfers does indeed reduce the probability of remarriage over two years, and that the elasticity of short-term remarriage with respect to the *AFDC* guarantee lies between 0 and $-.8$.

Though these results support contentions that the *AFDC* program contributes to growth in female headship, two important policy related questions remain unanswered. First, to what extent are the reduced remarriage probabilities caused by increased duration of search? It is conceivable that by increasing the duration of search, *AFDC* transfers permit female heads to locate "better" husbands than would be possible without such transfers. The argument parallels the claim that unemployment insurance promotes search and thereby leads recipients to find

---

1968 model may be written:

$$
\begin{aligned}
& \underset{(.7)}{2.6379} + \underset{(1.1)}{.0055X_3} - \underset{(2.8)}{.1054X_5} - \underset{(.4)}{1.6735X_6} \\
& - \underset{(1.0)}{.0033X_7} + \underset{(3.0)}{.4322X_8} + \underset{(2.1)}{1.279X_9} - \underset{(.1)}{.0258X_{10}} \\
& - \underset{(.5)}{.2371X_{11}} - \underset{(3.7)}{1.925X_{12}} + \underset{(.4)}{.2097X_{13}} + \underset{(.2)}{.0986X_{14}} \\
& + \underset{(.5)}{.5205X_{15}} - \underset{(1.4)}{.6811X_{16}} + \underset{(4)}{.2295X_{17}} \\
& + \underset{(.8)}{.4987X_{18}} - \underset{(1.1)}{.7799X_{19}} + \underset{(1.4)}{1.0642X_{20}}
\end{aligned}
$$

t-statistics are in parentheses

[28] In this regard, the Michigan researchers report that in 1969 panel losses were greatest among nonwhites with low incomes and low levels of education. This could explain the highly significant positive sign on the wage coefficient (X_8) in the 1968 model. Low-wage women who remarried may have been more likely to leave the sample.

higher wage jobs. (See Ronald Ehrenberg and Ronald Oaxaca.) In this case *AFDC* transfers could lead to improved marital stability and better homes for the children. Both results are consistent with the program's primary goals. Clearly, we need to know more about the relationship between *AFDC* transfers and the duration and productivity of marital search. A second important policy question concerns the extent to which alternative transfer programs that are available to both married and single individuals (such as negative income taxes, demogrants, and children's allowances) affect remarriage probabilities. Once again, such transfers may increase the duration of search and therby reduce short-term remarriage probabilities. The negative income tax experiments could shed new light on this subject.

REFERENCES

G. Becker, "A Theory of Marriage: Part I," *J. Polit. Econ.*, July/Aug. 1973, *81*, 813–46; "Part II," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, S11–26.

——, **E. Landes, and R. Michael,** "An Economic Analysis of Marital Instability," *J. Polit. Econ.*, Dec. 1977, *85*, 1149–87.

K. Burdett, "Theories of Search in a Labor Market," tech. analysis paper no. 13, Office of Evaluation, ASPER, U.S. Department of Labor, Washington 1973.

G. Duncan, "Unmarried Heads of Households and Marriage," in Greg J. Duncan and James N. Morgan, eds., *Five Thousand American Families*, Vol. 4, Ann Arbor 1976, 77–115.

R. Ehrenberg and R. Oaxaca, "Unemployment Insurance, Duration of Unemployment, and Subsequent Wage Gain," *Amer. Econ. Rev.*, Dec. 1976, *66*, 754–66.

A. Freiden, "The United States Marriage Market," *J. Polit. Econ.*, Mar./Apr. 1974, *82*, S34–53.

R. Gronau, "Information and Frictional Unemployment," *Amer. Econ. Rev.*, June 1971, *61*, 290–301.

M. Hannan, N. Tuma, and L. Groeneveld, "The Impact of Income Maintenance on the Making and Breaking of Marital Unions: Interim Report," Stanford Res. Instit., June 1976.

M. Honig, "AFDC Income, Recipient Rates, and Family Dissolution," *J. Hum. Resources*, Summer 1974, *11*, 303–23.

R. Hutchens, "Changes in AFDC Tax Rates: 1967–1971," *J. Hum. Resources*, Winter 1978, *13*, 60–74.

——, "State Policy Parameters and Recipient Behavior in the Aid to Families with Dependent Children Transfer System," unpublished doctoral dissertation, Univ. Wisconsin 1976.

S. Lippman and J. McCall, "The Economics of Job Search: A Survey," *Econ. Inquiry*, June 1976, *14*, 155–89.

D. Mortensen, "Job Search, the Duration of Unemployment, and the Phillips Curve," *Amer. Econ. Rev.*, Dec. 1970, *60*, 847–62.

M. Nerlove and S. J. Press, "Univariate and Multivariate Log-Linear and Logistic Models," R-1306-EDA/NIH, Rand Corp., 1973.

Heather Ross and Isabel Sawhill, *Time of Transition*, Washington 1975.

Isabel Sawhill et al., *Income Transfers and Family Structure*, Washington 1975.

J. Sweet, "Differentials in Remarriage Probabilities," unpublished paper, Center Demography Ecology, Univ. Wisconsin 1973.

Institute for Social Research, (*OEO*) *Panel Study of Income Dynamics*: *Study Design, Procedures*, Available Data *1968–72 Interviewing Years*, Vol. 1, Ann Arbor 1972.

Layoffs and Alternatives under Trade Unions in U.S. Manufacturing

By JAMES L. MEDOFF*

Firms faced with declines in the demand for their products choose among several forms of labor adjustment: reductions in employment; in average hours worked; and in (the growth of) hourly employee compensation. This study investigates how their choice is conditioned by trade unionism in the *U.S.* manufacturing sector. The investigation relies on evidence drawn from two sources: 1) various data sets containing information on industries, establishments, and individuals; and 2) major collective bargaining agreements.

Section I documents the primary finding that adjustment through layoffs is substantially greater in unionized firms than in comparable nonunionized firms. Section II argues that unions increase the use of layoffs by impeding the use of quits[1] and cuts in (the growth of) real wage rates, while restraining firms' ability to reduce average hours worked and discharge employees. The extent to which the union-nonunion differential in the use of layoffs is rewarded by a differential in unemployment insurance (*UI*)subsidy is examined in Section III.[2] Section IV considers the possibility that the role of layoffs is relatively large in unionized firms partially because under collective bargaining greater weight is attached to the preferences of senior workers, who are likely to prefer (inverse seniority-based) layoffs, than in a nonunion setting.[3] Section V summarizes the main results and their implications.

I. Layoffs

A. *Factors Influencing Layoff Rates*

"Layoffs" are defined by the Bureau of Labor Statistics (*BLS*) in its labor turnover reports (which are based on surveys of establishments) as "suspensions without pay lasting or expected to last more than 7 consecutive calendar days, initiated by the employer without prejudice to the worker." In the manufacturing sector during the 1958 to 1975 period, the mean monthly layoff rate (layoffs in month/persons on payroll in pay period including the twelfth day of month) was .017. By comparison, the mean monthly quit rate was .019 and the mean monthly other-separations rate was .008, about half of which is attributable to discharges.[4]

The layoff rate in a firm depends on several

*Assistant professor of economics, Harvard University; and National Bureau of Economic Research. I have benefitted from conversations with C. Brown, G. Chamberlain, M. Feldstein, R. Freeman, Z. Griliches, and R. Hall. K. Abraham, K. Clark, P Mueser, L. Summers, and the members of the Harvard Labor Seminar also made helpful suggestions that improved the quality of the research, as did participants in seminars at the University of Chicago, Cornell, and Washington University. I am very grateful to M. Van Denburgh for his skillful handling of all the computer work. I am also indebted to an anonymous referee for his insightful comments. The study was supported by the U.S. Department of Labor (Grant J-9-M-6-0094), the National Science Foundation (Grant APT77-16279), and the National Bureau of Economic Research (under its program of research on labor economics). It has not been reviewed by the Board of Directors of the National Bureau.

[1]The results concerning quits, derived with aggregate data, are consistent with the findings based on individual-level data presented by Richard Freeman (1978).

[2]For in-depth discussions of the *UI* system, see Joseph Becker (1972), Martin Feldstein (1974, 1975, 1976), and Martin Baily.

[3]For a general discussion of the determination of personnel policies in the presence of a union and in its absence, see Freeman (1976a), and Freeman and the author forthcoming.

[4]The definition of layoffs and an explanation of the employment concept used in calculating rates are found in *BLS* (1976a, pp. 776–77). The estimated mean rates were derived with information from *BLS* (1976a, pp. 41–42), and *BLS* (1977, p. 101); they are unweighted averages of annual mean monthly rates for the manufacturing sector as a whole. The statement concerning the relative importance of discharges is based on unpublished data provided by *BLS*, which are discussed at greater length below.

factors which determine the benefits and costs of suspending workers: changes in the level of product demand; the nature of the firm's technology; the likelihood that employees will return to the firm if recalled; and the feasibility of using other modes of adjustment to market changes, such as voluntary attrition, reductions in average hours worked, and cuts in (the growth of) real wage rates. In the empirical work, industry dummies and (in some regressions) the mean and variance of monthly differences between accessions and separations rates are supposed to capture shifts in demand that occur during a given period of years; establishment size, hours per employee, and the capital-labor ratio are viewed as controls for within-industry variation in technology; and the wage rate and measures of labor "quality" are seen as determinants of the probability that firms will be able to rehire laid-off workers. Unionism is expected to raise layoff rates by restricting the use of other adjustment mechanisms, for reasons to be discussed later.

Because the layoff rate in state by industry or industry cell j (l_j) is a fraction varying between 0 and 1, it was represented by a logistic equation:

$$(1) \qquad l_j = \frac{e^{z_j'\beta}}{1 + e^{z_j'\beta}}$$

where the vector β contains the parameters to be estimated and the vector z_j represents the factors mentioned above influencing the layoff rate in j. Equation (1) can be rewritten in the "*log* odds ratio" form:

$$(2) \qquad ln\left(\frac{l_j}{1 - l_j}\right) = z_j'\beta$$

which was used in the empirical analysis.

As Takeshi Amemiya and Fred Nold have pointed out, the standard logit model (equation (2)) should be modified to include an equation error, which can be interpreted as a surrogate for omitted variables. An additional issue discussed in Joseph Berkson (1944, 1953, 1955) is that since frequencies are used in constructing the *log* odds ratio, the logit equation will have heteroscedastic residuals. If these frequencies are based on independent samples from binomial populations, the estimator of the true *log* odds ratio will have an asymptotic variance which is approximately equal to ω_j, defined as $\omega_j = \sigma^2 + [N_j l_j(1 - l_j)]^{-1}$, where σ^2 is the variance of the equation error and N_j is the number of observations in sample j.[5] The variance σ^2 can be estimated consistently as $s^2 = (1/T)\{\Sigma_j(e_j - z_j'\mathbf{b})^2 - \Sigma_j[1/(N_j l_j(1 - l_j))]\}$, where T is the number of cells, e_j is the *log* odds ratio for cell j, and $\mathbf{b}$ is a vector of logistic parameters estimated by ordinary least squares (*OLS*).

In an effort to obtain efficient parameter estimates and consistent standard errors, each logistic turnover rate model to be discussed was fit by weighted least squares (*WLS*). The value of $(\omega)^{-1/2}$ used to multiply each variable prior to least squares estimation was calculated with N_j equal to the number of employees in all establishments or in sampled establishments (depending on the data set) in cell j.[6]

B. *Empirical Analyses of Layoff Rates*

The layoff rate model set out above was estimated with two data files for the manufacturing sector: a state 2-digit SIC industry file for 1965–69 and a 3-digit SIC industry file for 1958–71.

1. *Estimates Based on State 2-Digit SIC Data*

The state 2-digit SIC data came from several sources. Turnover rates for 1965–69 were obtained from a computer tape provided by *BLS*. Information for the thirty states which reported 2-digit SIC industry statewide rates in each year of the period under analysis

[5]This result is derived in Amemiya and Nold; see also Berkson (1944, 1953, 1955) and Henri Theil.

[6]With state 2-digit Standard Industrial Classification (SIC) data for 1965 to 1969, N_j was set equal to the average number of employees in cell j during these years, since no information on the number in sampled establishments was available. With 3-digit SIC data for 1958 (or 1959) to 1971, N_j was set equal to the average number of employees in sampled establishments in industry j from 1963 to 1971, the only years for which *BLS* provided this information. If it is assumed that $\sigma^2 = 0$, $(\omega)^{-1/2}$ reduces to the Berkson *WLS* correction factor.

was used. An unweighted average of annual mean monthly layoff rates was calculated to give the frequency employed in constructing the *log* odds ratio dependent variable. Monthly differences between accessions and separations rates were also computed; the unweighted mean and variance of these differences were included in some regressions, since better proxies are not available for that portion of product demand variation uncaptured by industry and region dummies.

Estimates of the fraction of workers who are union members and an index of labor "quality" were derived on a Current Population Survey (*CPS*) state group-by-industry basis with data from the 1973, 1974, and 1975 May *CPS* files. The fraction unionized variable was based on weighted (with the *CPS* sampling weights) counts of private sector wage and salary labor force members. For comparability with the *CPS* unionization estimates, the *CPS* classification of states into twenty-nine groups was adopted in constructing all other variables.[7] The derivation of the labor quality index is described in detail in Charles Brown and the author. Briefly put, it was calculated by first imputing to each private wage and salary worker in a state-by-industry cell a predicted usual hourly wage that reflects the individual's age, schooling, region of residence, and sex (but not his or her union membership status), and then using the *CPS* sampling weights multiplied by usual weekly hours to derive a weighted average. The predicted wages came from *ln* (real usual hourly wage) functions fit separately for workers of each sex with the *CPS* micro data. These regressions included a union membership dummy, in addition to the other personal characteristics used in deriving the predicted wages.

Average hourly wages, average hours per employee, and the average capital-labor hours ratio for 1965–69 were calculated for each state by industry cell with data from various editions of the *Annual Survey of Manufactures* and the *1967 Census of Manufactures.* The wage data from these sources were inflated to 1972 dollars with the Consumer Price Index (*CPI*). The average hours estimates were based on the assumption that nonproduction employees (for whom no hours data are provided) worked 2,080 hours per year. The procedure used for estimating each cell's capital stock in 1965–69 is described in Brown and the author; it involves adding gross investment flows to a 1964 estimate of the gross book value of capital, and making an adjustment for capital that is rented. The *Annual Survey of Manufactures* does not present data on the number of establishments in each state by industry cell, but the *Census of Manufactures* does. Thus, only information for 1967 on the number of establishments in each cell could be used in creating a labor hours per establishment variable.

Only cells in which there were at least twenty-five private wage and salary *CPS* sample members with all of the information required to calculate the unionization and labor quality index variables were used in the regressions presented below. Cells also had to pass the tests of data availability discussed in Brown and the author. As a result of applying these inclusion criteria, only nineteen out of twenty-one 2-digit SIC industries are represented in the analysis.

Table 1 presents the estimates of the impact of unionism on layoff rates based on the state 2-digit SIC data. Regression 1 implies that in the 1965–69 period, workers in unionized establishments had a substantially and significantly higher probability of being laid off than did workers in similar nonunion firms. The estimated coefficients from this regression and the mean values for the whole sample of the independent variables (other than fraction unionized) can be plugged into the logistic function (equation (1)) to provide estimates of the average monthly probability of a layoff in union and nonunion establishments with characteristics the same as the total sample means. The estimated average monthly layoff rate from 1965 to 1969 for nonunion establishments is .005; for similar unionized establishments, the comparable estimate is .023.

The mean and the variance of monthly differences in accessions and separations rates are included in regression 2. Inclusion of these two crude proxies reduces the estimated coef-

[7]Thus, the actual unit of observation is a *CPS* state group (as opposed to state) 2-digit SIC industry cell.

TABLE 1—LAYOFF RATE REGRESSIONS: STATE 2-DIGIT SIC MANUFACTURING DATA FOR 1965–69[a] ($\bar{N}$ = 151)

	Dependent Variable: *Ln* (Layoff Rate/(1-Layoff Rate))[b]		
Coefficients of:	Unweighted Mean [*S.D.*]	*WLS* Regression Numbers:[c] 1	2
Fraction Unionized	317	1.496	1.233
	[.186]	(.428)	(.416)
Mean Monthly Difference between Accessions and Separations Rates (hundredths)	.146	–	−.311
	[.398]		(.178)
Variance of Monthly Differences between Accessions and Separations Rates (hundredths)	.106	–	.097
	[.820]		(.082)
North East Dummy	.255	.043	−.036
	[419]	(.149)	(.145)
South Dummy	.384	−.048	−.033
	[.488]	(.185)	(.178)
West Dummy	.073	.304	.165
	[.261]	(.225)	(.228)
Industry Dummies (19)	–	yes	yes
Mean Hourly Wage (1972 $)	3 871	−.020	−.009
	[.721]	(205)	(.196)
Labor "Quality" Index	3 523	021	.120
	[.444]	(.257)	(.248)
Average Hours per Employee (thousands)	2.032	−2 885	−2.277
	[073]	(1.357)	(1.308)
Labor Hours per Establishment (millions)	.210	−1 443	−1.341
	[.222]	(.354)	(.344)
Mean Capital-Labor-Hours Ratio (historical $)	6 464	.031	.030
	[4.268]	(0.27)	(.026)
Standard Error of Estimate[d]	–	.583	.560

Note: Standard errors are in parentheses.

[a]Unweighted Mean of Layoff Rate = .012 [*S.D* = .018]

[b]Unweighted Mean of Dependent Variable = −4.778 [*S D* = .841]

[c]The weighting procedure used is described in Section IA.

[d]This statistic was estimated as $[\Sigma_i(v_i - v^*)^2/(N - K)]^{1/2}$ where the v_i are residuals derived with the unweighted values of variables (and the *WLS* parameter estimates), v^* is the mean of the v_i, N is the number of observations, and K is the number of independent variables.

ficient of the unionization variable. Nevertheless, the union coefficient estimate remains large and statistically significant despite the fact that it is most likely biased downwards (because the variable giving the mean monthly difference between accessions and separations rates, which has a negative partial correlation with the unionization variable, has an estimated coefficient that is most likely biased downwards due to that proxy's definitional relationship to the mean monthly layoff rate).[8]

2. *Estimates Based on 3-Digit SIC Data*

The 1958–71 data for 3-digit SIC industries used to fit the layoff rate model were assembled as follows.[9] Unweighted averages of annual mean monthly layoff rates were derived from a computer tape version of

[8]The results in Table 1 are nearly identical to those obtained using *OLS*. The *OLS*-estimated fraction unionized coefficient (standard error) was 1.506 (.435) in model 1 and 1.245 (.426) in model 2. The Table 1 estimates differed more from those derived when Berkson weights (see fn. 6) were used. With the Berkson weighting, the estimated coefficient (standard error) of the fraction unionized variable was 1.992 (.514) in model 1 and 1.040 (.429) in model 2.

[9]To merge the various data sets employed, it was necessary to aggregate some of the 3-digit SIC industries.

Employment and Earnings: United States, 1909–1975. This tape also provided information on the sex composition of the work force and on monthly accessions and separations rates.

The fraction of all employees covered by collective bargaining agreements was estimated with data from *BLS*'s 1968, 1970, and 1972 Expenditures for Employee Compensation (*EEC*)[10] establishment-level files, supplemented by industry-level information (on establishments excluded from the publicly available tapes to preserve confidentiality) provided by *BLS*. A more detailed description of these estimates is found in Freeman and the author (1979). The *EEC* files were also used to approximate the fraction of employees in establishments contributing to Supplemental Unemployment Benefits (*SUB*) and/or severance pay funds, and employer contributions to these funds.[11]

Data for 1958–69 on total wages, production and nonproduction worker employment, and production worker hours came from *Industry Profiles 1958–1969*. The same data for 1970 and 1971 came from *Annual Survey of Manufactures 1970–71*. The wage data were inflated to 1972 dollars with the *CPI*. Estimates of total hours worked were derived under the assumption that nonproduction employees worked 2,080 hours per year. For reasons of feasibility and classification comparability, establishment counts for 1967 from the *1967 Census of Manufactures* were used in computing a labor hours per establishment variable.

Data for 1958–71 on privately owned capital came from a tape furnished by *BLS*. The *BLS* estimates were calculated using a perpetual inventory method that accumulated deflated plant and equipment investment flows adjusted to reflect depreciation. A more detailed description of the derivation of these capital stock figures can be found in Jack Faucett Associates, Inc.

Table 2 presents layoff rate regressions fit with the 3-digit SIC manufacturing data.[12] The estimated coefficients in regression 1 imply that the average monthly layoff rate during the 1958–71 period was .010 in nonunion firms with characteristics the same as the total sample means and .022 in similar ones covered by collective bargaining. When the mean and variance of monthly differences between accessions and separations rates are included (regression 2), the estimated coefficient of the collective bargaining coverage variable is lower but remains substantial, despite the fact that it is likely to be biased downwards (for the reason given in discussing the state-by-industry results).

Benefits from private unemployment insurance funds, under plans that cover the workers in only one firm, do not represent a layoff-related subsidy (unlike imperfectly experience-rated *UI* benefits).[13] However, because of imperfections in the capital market, they can affect a firm's layoff rate by reducing the expected cost to employees of being laid off. To examine the importance of *SUB* and/or severance pay funds as mediating factors in the estimated union-layoff rate association, variables equal to the fraction of employees in establishments contributing to these funds and employer contributions to them were included in regression 3. While entering these two variables does not lead to a substantial reduction in the estimated coefficient of the fraction covered by collective bargaining, this estimate is likely to be biased in an unknown direction due to endogeneity and mismeasurement of the *SUB* and/or severance pay funds variables.[14]

[10]The *EEC* survey is described in detail in *BLS* (1976b, pp. 175–83). Sampling weights were used in the computation of industry-level variables (but not the regressions) with the *EEC* files.

[11]Unfortunately, information about each fund by itself is not collected. Due to the fact that *BLS* withheld the survey responses of a small number of very large potentially identifiable establishments, these two variables are likely to be quite mismeasured. The pre-1972 figures on contributions to funds were inflated to 1972 dollars with the *CPI*.

[12]The appropriate *F*-test indicates that the set of 2-digit SIC industry dummies is statistically significant at better than the .05 level in each of the Table 2 regressions. When no industry dummies were included, the estimated coefficient of the fraction covered by collective bargaining variable was always substantially larger than in the comparable Table 2 regression.

[13]See Bureau of Labor Statistics (1965) and Becker (1968) for discussion of the history and functioning of private unemployment insurance plans.

[14]The results in Table 2 are nearly identical to those obtained using *OLS*. The *OLS*-estimated coverage coeffi-

TABLE 2—LAYOFF RATE REGRESSIONS: 3-DIGIT SIC MANUFACTURING DATA FOR 1958–71[a]
(N = 89)

	Dependent Variable: *Ln* (Layoff Rate/(1-Layoff Rate))[b]			
Coefficients of:	Unweighted Mean [*S.D.*]	*WLS* Regression Numbers:[c] 1	2	3
Fraction of Employees Covered by Collective Bargaining	.459 [.192]	.841 (.320)	.808 (.282)	.771 (.289)
Mean Monthly Difference between Accessions and Separations Rates (hundredths)	−.106 [.143]	–	−.917 (.359)	−.604 (.410)
Variance of Monthly Differences between Accessions and Separations Rates (hundredths)	037 [.083]	–	1.680 (.611)	1.796 (.615)
Industry Dummies (21)	–	yes	yes	yes
Mean Hourly Wage (1972 $)	3.855 [.711]	.046 (203)	.045 (.180)	−.028 (.186)
Fraction Male Workers	.752 [.175]	− 339 (.667)	−.561 (.585)	−.375 (.597)
Fraction Production Workers	.768 [096]	2.013 (1.001)	1.264 (.898)	1.375 (.915)
Average Hours per Employee (thousands)	2 018 [068]	−2.785 (1 239)	−2.015 (1.159)	−1.519 (1.207)
Labor Hours per Establishment (millions)	.291 [.344]	−.636 (.233)	− 597 (.195)	− 696 (.205)
Mean Capital-Labor-Hours Ratio (1972 $)	7.227 [7.422]	−.007 (014)	− 012 (.013)	−.015 (.013)
Fraction in Establishments Contributing to *SUB* and/or Severance Pay Funds	077 [139]	–	–	1.027 (931)
Per Employee "Employer" Contributions to *SUB* and/or Severance Pay Funds (1972 $)	7.736 [21.475]	–	–	−.002 (.005)
Standard Error of Estimate[d]	–	.384	.336	.335

Note: Standard errors are enclosed in parentheses.
[a]Unweighted Mean of Layoff Rate = .017 [*S D* = 011]
[b]Unweighted Mean of Dependent Variable = −4.253 [*S D* = .601]
[c]The weighting procedure used is described in Section IA.
[d]See fn. d in Table 1

C. *On-Layoff Frequencies*

One criticism of the preceding analysis might take the following form: the estimates of the union-nonunion layoff rate differential presented above only reflect the fact that the component industries in each 2-digit SIC grouping which have higher than average product demand variation (not captured by the proxies used), also tend to have higher than average unionization. If this were the case, the estimated layoff rate differential would be expected to be much closer to zero within much more disaggregated industries, since differences in the product demand variation faced by union and nonunion establishments would be expected to be much closer to zero.

Another criticism might be formulated as follows: the estimates of the union-nonunion layoff rate differential only reflect the fact that the average duration of a completed spell on layoff is shorter for workers laid off from unionized firms.[15] If this argument were

cient (standard error) was .837 (.319) in model 1, .805 (281) in model 2, and .768 (.288) in model 3. The Table 2 estimates differed more from those derived when Berkson weights (see fn. 6) were used. With the Berkson weighting, the estimated coverage coefficient (standard error) was .903 (.405) in model 1, .701 (.382) in model 2, and .722 (.374) in model 3.

[15]Regressions of *ln* (weeks on layoff) on industry dummies and a union member dummy fit with data for 839 private wage and salary blue-collar manufacturing workers on layoff as of the 1973, 1974, or 1975 May *CPS* surveys indicate that the duration of unemployment spells as of the relevant surveys was somewhat shorter for union members. The estimated coefficient (standard

correct, it would be expected that within industries the fraction of union members on layoff at a given point in time would on average not be greater than the fraction of nonmembers, since these fractions reflect both layoff rates and average durations of completed spells on layoff.

To examine these possibilities, a merged file of the 1973, 1974, and 1975 May *CPS* was analyzed. Unlike the *BLS* turnover rate surveys which elicit information on an establishment's layoff rate (as defined in Section IA), the *CPS* surveys obtain information on whether or not an individual was absent from work in the week prior to the survey because he or she was on "Temporary layoff (under 30 days)" or "Indefinite layoff (30 days or more or no def. [definite] recall date)." Individuals on temporary or on indefinite layoff are aggregated into one group (as in Feldstein, 1975) for the subsequent analysis and classified as "on layoff."[16]

On-layoff frequencies for blue-collar private wage and salary workers grouped by 2- or 3-digit Census industry were derived with the *CPS* data. An industry was included in the analysis if at least one union blue-collar worker, at least one nonunion blue-collar worker, and at least one blue-collar worker on layoff from the industry were surveyed in May of 1973, 1974, or 1975. (Twenty-one out of twenty-one 2-digit Census industries (with 26,326 sample members) and seventy-two out of eighty-three 3-digit Census industries (with 25,851 sample members) were included; it turned out that every industry that passed the inclusion criteria had at least 12 union and 15 nonunion blue-collar sample members.) Each on-layoff frequency was calculated as the ratio of blue-collar sample members on layoff to blue-collar sample members employed or on layoff. The within-

error) of the union member dummy (member = 1) was −.153 (.079) when the Census industry dummies were at the 2-digit level and −.156 (.082) when they were at the 3-digit level.

[16]It should be noted that the *CPS* on-layoff frequencies, unlike the state by 2-digit SIC and the 3-digit SIC layoff rates, do not reflect individuals whose layoff is likely to be a permanent separation, but do reflect individuals on layoff for one week or less.

TABLE 3—WITHIN-INDUSTRY DIFFERENCES BETWEEN UNION AND NONUNION ON-LAYOFF FREQUENCIES FOR BLUE-COLLAR WORKERS: BASED ON 1973–75 MAY *CPS* DATA

	2-Digit Census Industries	3-Digit Census Industries
Industries where Union On-Layoff Frequency was Greater/Total Number of Industries	16/21	46/72
Probability that an Equal or Greater Fraction Would be Observed if *Pr* (union frequency > nonunion frequency) = .5	.015	.013
Logistic Estimate of Amount by which Union On-Layoff Frequency Exceeded Nonunion Frequency in the Same Industry[a]	.0119	.0114

[a]The derivation of these estimates is discussed in fn 17.

industry differences between union and nonunion on-layoff frequencies are summarized in Table 3.

The numbers in the table demonstrate that blue-collar workers who were union members had a significantly and substantially higher probability of being on layoff than did nonmembers within the same 2- or 3-digit Census industry. In addition, they show that the union-nonunion differential in on-layoff frequencies within seventy-two 3-digit industries was nearly the same as the differential within twenty-one 2-digit industries.[17] Other

[17]The differentials presented in Table 3 are based on *WLS* regressions in which the *log* odds ratio of the on-layoff frequency in an industry by union member status cell was regressed on a union member and industry dummy variables. The weighting scheme used is described in Section IA; N_j was set equal to the number of sample members from cell j who were employed or on layoff, and since $s^2 < 0$ (for both of the models estimated), it was assumed that $\sigma^2 = 0$. (This implies that the weights employed are the traditional Berkson weights.) In light of the discussion in Berkson (1953, 1955), if cell j had a zero on-layoff frequency (none did at the 2-digit Census industry level, but 13 percent did at the 3-digit level), it was assigned a frequency equal to $1/2N_j$. The estimated coefficient (standard error) of the union member dummy (member = 1) was .441 (.090) when the Census industrial classification was 2-digit and

experiments with the *CPS* micro data indicate that controlling for age, sex, race, schooling, *CPS* state group of residence, usual weekly hours, and usual weekly earnings does not alter the basic conclusions of this analysis.

II. Alternatives to Layoffs and Implications for Accessions

The previous section has provided evidence that labor adjustment through layoffs is much more important in manufacturing establishments that are covered by collective bargaining than in comparable ones that are not. This section argues that unions increase the use of layoffs by lowering quits and reducing the responsiveness of (the growth of) real wage rates to demand conditions, while limiting the extent to which average hours can be reduced and employees can be discharged. In addition, it will be shown that the greater use of layoffs in unionized establishments during periods of decreased commodity demand permits a higher rate of rehires and lower rate of new hires when commodity demand increases. Thus, the section contends that the entire labor adjustment process is very different in the presence of collective bargaining.

A. *Quits*

Freeman's paper (1978) provides strong support for the claim that unionization brings about a large reduction in quit probabilities, even when wages and fringe benefits are held constant. To obtain results comparable with the layoff rate findings, I estimated *log* odds ratio quit rate models of exactly the same form as 1 and 2 in Tables 1 and 2 (note that each model includes a mean hourly wage variable).[18] With the state 2-digit SIC data set, the estimated coefficient (standard error) of the fraction unionized variable was −.447 (.192) in model 1 and −.451 (.184) in model 2. With the 3-digit SIC data, the collective bargaining variable had an estimated coefficient (standard error) equal to −.225 (.126) in model 1 and −.185 (.122) in model 2.[19] (The unweighted monthly quit rate mean [standard deviation] is .027 [.012] with the state 2-digit SIC file and .019 [.008] with the 3-digit SIC file.) Thus, the ability of a firm to adjust its employment through voluntary attrition appears to be reduced substantially by the net benefits workers receive under unionism.

B. *Wage Rate (Growth) Reductions*

Another alternative to layoffs that, in theory, is available to management is a reduction in (the growth of) real hourly employee compensation. The fact that the relative wage impact of unionism has tended to be greater during economic downturns has been observed by a number of individuals, in particular by H. Gregg Lewis. This observation suggests that the reduction in (the growth of) real wage rates in response to a reduction in product demand is smaller under trade unions.

Between May 1973 and May 1975, when the percent of experienced private sector wage and salary workers in manufacturing who were unemployed rose from 4.3 to 11.7, the expected increase in the union relative wage effect occurred. This statement is supported by regressions of a *ln* (hourly wage rate) variable on a union member dummy (member = 1), 3-digit Census industry dummies, and variables representing numerous personal characteristics, fit with May 1973 and May 1975 *CPS* data for private sector blue-collar

385 (.085) when it was 3-digit. (The mean on-layoff frequency was .032 for individuals in the 2-digit sample and .033 for those in the 3-digit sample.)

[18]Unweighted averages of annual mean monthly quit rates were employed in constructing the *log* odds ratio dependent variables. The 1965–69 state 2-digit SIC and the 1958–71 3-digit SIC quit and new hires rates used in Section II came from the *BLS* tapes from which the layoff rates were obtained.

[19]Under *OLS*, the estimated unionization coefficient (standard error) with the state 2-digit SIC data was − 441 (.191) in model 1 and −.443 (.184) in model 2, and with the 3-digit SIC data −.221 (.125) in 1 and −.181 (.121) in 2. Using Berkson weights (see fn. 6), the estimate with the state 2-digit data was −.383 (.210) in 1 and −.459 (.193) in 2, and with the 3-digit data −.361 (.147) in 1 and −.273 (.137) in 2.

manufacturing workers paid by the hour.[20] The estimated coefficient (standard error) of the union member dummy was .160 (.007) for May 1973 and .181 (.008) for May 1975. (This increase is significant at the .05 level.)

It should be noted that since roughly 3.8 million union members in manufacturing were covered by cost-of-living adjustment clauses as of the beginning of 1975,[21] analyzing the May 1973–May 1975 period, in which there was rapid and increasing inflation in addition to a large increase in the rate of unemployment, may yield results that are the exception rather than the rule. However, the conclusion that the growth of real hourly wages is somewhat less responsive to demand conditions in the presence of a union is not inconsistent with the findings presented by Lewis (pp. 195–241) who, in a study of the 1920–58 period, controlled for changes in the general price level (which were negatively related to changes in the impact of unionism on relative wages).

The claim that in unionized firms there is a relatively strong preference for real wage (growth) rigidity over employment stability is consistent with the fact that most union-management agreements are long-duration contracts under which (a lower bound to) the growth of money wage rates during the contract period is agreed to before the period begins.[22] In 1974, 65 percent of the "major contract" (one covering 1,000 or more workers) manufacturing work force was covered by agreements whose duration was three years or more, nearly all of which had provisions for automatic money wage increases, and only 1 percent was covered by agreements whose duration was one year or less.[23]

[20]The personal characteristics variables included were a sex dummy, a race dummy, years of schooling completed, age − schooling − 5 and its square, three marital status dummies, number of dependents, three occupation dummies, and twenty-eight *CPS* state-group-of-residence dummies. The sample size was 5,794 for May 1973 and 4,817 for May 1975.

[21]This estimate is from Harry Douty, p. 18.

[22]This statement is based on data found in *BLS* (1975a, pp. 7, 40).

[23]See fn. 22.

C. *Average Hour Reductions*

With lower quit rates and less ability to reduce (the growth of) real wage rates, unionized firms must make greater use of other adjustment mechanisms, such as layoffs, average hour reductions, and discharges. Trade unions are by no means indifferent as to the relative importance of each of the possibilities.

At an earlier point in time many unions wanted a moderate amount of worksharing to occur before layoffs were initiated. By 1960, this had changed according to Sumner Slichter, James Healy, and E. Robert Livernash:

> Whereas twenty years ago a substantial number of unions insisted on reasonable work-sharing before layoffs could be made, today the trend of union preference is more and more toward the restriction of work-sharing arrangements. Unions have always opposed work-sharing that was carried to the extent of "sharing misery," but the restrictions sought by unions in 1958 and 1959 were intended to go far beyond the usual controls. In some cases the union asked that layoffs be used exclusively without any work-sharing. [p. 152]

Comparison of the nature of hour-reduction provisions in major contracts in effect in 1954–55 with those in effect in 1970–71 strongly suggests growth in the union preference for layoffs over work sharing. In 1954–55, only 5 percent of the 28 percent of the major contract work force covered by hour-reduction provisions had agreements stating that layoff proceedings would begin when hours worked were below normal for four or less weeks. By 1970–71, a dramatic change had occurred: 43 percent of the 28 percent covered by hour-reduction provisions had contracts under which layoff proceedings would commence after four or less short work weeks.[24] In 1954–55, 31 percent of the workers covered by reduction-in-hours provisions had negotiated a guarantee of union partici-

[24]The data on the major 1954–55 contracts are from *BLS* (1957, pp. 2, 8, 9); the 1970–71 agreement data are from *BLS* (1972a, pp. 22, 24).

pation in the choice between reduced hours and immediate layoffs. By 1970–71, the comparable figure had increased sharply to 72 percent.[25] According to the *BLS* one of the main reasons that a union wants a meaningful voice in decisions involving work sharing is that "if... it is known that manpower needs will be curtailed for a lengthy period, the union may prefer to bypass the reduced hour provisions and initiate layoffs immediately" (1972a, p. 16).

The sharp downturn from May 1973–May 1975 provides evidence consistent with the claim that a reduction in the average weekly hours of those working is a less acceptable substitute for layoffs under trade unions. Data from the May 1973 and May 1975 *CPS* surveys for private sector blue-collar manufacturing workers paid by the hour were used to calculate the change in the on-layoff frequency and the change in the natural logarithm of average hours worked (for those working) in the week prior to the relevant surveys for union members and nonmembers within 3-digit Census industries. For union members, the cross-industry weighted average change in the on-layoff frequency was .064 and the weighted average change in the natural logarithm of average hours worked was −.044. For nonmembers, the comparable weighted averages were .031 and −.034.[26] Thus, the total adjustment through layoffs and per employee hour reductions was on average greater in the unionized sector of 3-digit Census industries, as would be expected given the impact of unionism on quits and real wage (growth) downward flexibility. Moreover, while adjustment through layoffs was 45 percent more important than adjustment through a reduction in average hours for hourly blue-collar union members, adjustment through layoffs was 9 percent less important for hourly blue-collar nonunion employees, which is consistent with a stronger preference for layoffs over work sharing under collective bargaining.

D. *Discharges*

The amount of labor input employed by a firm is also a function of its rate of other separations. The other separations category includes discharges, "terminations of employment initiated by the employer for such reasons as incompetence, violation of rules, dishonesty, laziness, absenteeism, insubordination, failure to pass probationary period, etc."(*BLS*, 1976b, p. 44), retirements, separations due to permanent disability, deaths, transfers to another establishment of the company, and entrances into the Armed Forces. Unpublished data provided on tape by *BLS* indicate that the unweighted average of 1959–71 mean monthly discharge rates for the manufacturing sector was .0038, approximately 50 percent of the comparable other separations rate. Retirements most likely were responsible for more than half of the remainder, with interplant transfers next in importance.

Discharge rate regressions (using unweighted averages of 1959–71 mean monthly rates) of exactly the same form as layoff rate regressions 1 and 2 in Table 2 were fit. In model 1, the estimated coefficient (standard error) of the fraction covered by collective bargaining variable was .023 (.175). When the mean and variance of monthly differences between accessions and separations rates were added, the estimated coefficient (standard error) was .038 (.171).[27] Thus, although the coefficient estimates are quite imprecise, it appears that the discharge rate, unlike the

[25]The 1954–55 and 1970–71 union participation estimates are from *BLS* (1957, p. 14); (1972a, p 15).

[26]Each of the weighted averages was based on the seventy-six industries for which all of the requisite information could be obtained The weight given to industry j was $N_jU_j(1-U_j)/\Sigma_jN_jU_j(1-U_j)$, where N_j is the number of employed or on-layoff hourly blue-collar sample members from industry j surveyed in May 1973, and U_j is the fraction of these workers who were union members. These weights were chosen since they implicitly weight within-industry union effects in an *OLS* regression with individual-level data of a variable on a union member dummy and industry dummies. Weighted averages were also calculated with the jth industry's weight equal to N_j/Σ_jN_j; they were nearly identical to the ones presented.

[27]Under *OLS*, the estimated coverage coefficient (standard error) was .047 (.173) in model 1 and .066 (.168) in model 2. Using Berkson weights (see fn. 6), the estimate was −.130 (.189) in 1 and −.110 (.191) in 2.

layoff rate, is not more than trivially higher under trade unions.[28]

E. *Implications for Accessions*

It has been shown above that trade unionism has a substantial effect on the mix of a firm's separations. As a result, it would be expected that the relative importance of new hires and rehires (or recalls) would be substantially different in establishments that are unionized than in comparable ones that are not. Unfortunately, *BLS* only collected data on total accessions, which include interplant transfers, and new hires prior to 1976. Thus, a precise estimate of the rehires rate for the period under analysis does not exist. However, the available information on the composition of the other separations category suggests that from 1959 to 1971 rehires represented roughly 85 percent of the difference between total accessions and new hires. In what follows, rehires will be measured by this difference.

Table 4 presents *OLS* and *WLS* regressions in which the dependent variable is *ln* (unweighted average monthly rehires rate ÷ unweighted average monthly new hires rate) and the independent variables are the same as in regression 1 or 2 in Tables 1 and 2. For the *WLS* estimation, each variable was multiplied by $(N_j)^{1/2}$, where N_j is the number of employees in all establishments (with the state 2-digit SIC file) or in sampled establishments (with the 3-digit SIC file) in cell j.[29] With the state by industry data, the *OLS* and *WLS* results indicate that the ratio of rehires to new hires is very much greater under trade unions. With the 3-digit SIC data, the estimated differentials are lower, but are still very large. Thus, employment adjustments in similar union and nonunion manufacturing establishments take very different forms: under trade unions, quits and new hires are much less important and layoffs and rehires much more important.

III. The Union-Nonunion Differential in *UI* Subsidy

Unemployment insurance benefits are neither perfectly experience rated nor subject to payroll or income taxation. As a result, there is an incentive for a firm to choose layoffs over reductions in average hours or (growth in) real wage rates. (See Feldstein, 1976; Baily.) This section examines the union-nonunion differentials in per employee *UI* taxes and benefits. The results provide additional confirmation of the claim that labor adjustment through layoffs is greater under unionism. In addition, they indicate roughly the extent to which the differential in wages lost through time on layoff is offset by the operation of the *UI* system.

The union-nonunion differential in *UI* subsidy can be analyzed as follows. Let s represent an annual *UI* subsidy ($s > 0$) or loss ($s < 0$) per employee, α an average *UI* benefit per laid-off worker, l an annual layoff rate, and τ an amount of *UI* taxes paid per employee per year. To simplify the analysis at what appears to be little cost, assume that α is the same for individuals who are covered by a collective bargaining agreement and for those who are not.[30] Under this assumption, the

[28] Similar regressions with a *log* odds ratio dependent variable based on differences between other separations and discharge rates were also fit. The estimates of the fraction covered coefficients, while imprecise, suggest that the monthly retirement rate from 1959–71 was somewhat higher under unionism. However, given the likely magnitude of monthly retirement rates, a union-nonunion difference is likely to be quite small compared to, say, the absolute value of the difference in monthly quit rates. Nevertheless, the role of retirements as an alternative to layoffs under trade unions should be explored with more refined data in future research. For related discussions, see Freeman (1976), E. Hodgens, and J. Zalusky.

[29] See fn. 6 for a related discussion. This weighting represents an attempt to provide efficient parameter estimates and consistent standard errors.

[30] While the average completed spell on layoff is likely to be somewhat shorter for union members than for nonmembers in the same industry (given the results presented in fn. 15 on the difference in average durations on layoff as of the *CPS* surveys analyzed), the average weekly benefits are likely to be somewhat higher for the former group as a result of the union wage effect (and in spite of the maximum weekly benefit amount specified under each state system).

TABLE 4—REHIRES/NEW HIRES REGRESSIONS: STATE 2-DIGIT SIC AND 3-DIGIT SIC MANUFACTURING DATA[a]

	Dependent Variable: *Ln* (Rehires Rate/New Hires Rate)[b]							
	State 2-Digit SIC; 1965–69 (N = 151)				3-Digit SIC; 1958–71 (N = 89)			
	Regression Numbers:[c]							
Coefficients of:	*OLS*1	*WLS*1	*OLS*2	*WLS*2	*OLS*1	*WLS*1	*OLS*2	*WLS*2
Fraction Unionized or Covered	1.195 (.343)	1.104 (376)	1.196 (.352)	.938 (.379)	.820 (.238)	.713 (.293)	.789 (.218)	.589 (.265)
Mean Monthly Difference between Accessions and Separations Rates (hundredths)	–	–	−.046 (.151)	−.397 (.165)	–	–	−.691 (.278)	−.632 (.306)
Variance of Monthly Differences between Accessions and Separations Rates (hundredths)	–	–	−.037 (.070)	−.149 (.069)	–	–	.981 (.473)	1.597 (.678)
Region Dummies (3)	yes	yes	yes	yes	no	no	no	no
Industry Dummies (19 or 21)	yes	yes	yes	yes	yes	yes	yes	yes
Mean Hourly Wage (1972 $)	306 (.164)	.480 (.156)	.305 (.166)	.410 (.156)	.549 (.152)	.625 (.161)	.540 (.140)	.568 (.149)
Labor "Quality" Index	.103 (.206)	−.131 (.222)	.108 (209)	−.092 (.220)	–	–	–	–
Fraction Male Workers	–	–	–	–	−.863 (.498)	−1 269 (.561)	−1.017 (.452)	−1.060 (.504)
Fraction Production Workers	–	–	–	–	3.159 (748)	4.149 (.711)	2.629 (.696)	3.430 (.667)
Average Hours per Employee (thousands)	.198 (1.085)	236 (1.101)	.138 (1.103)	− 038 (1 100)	−1.459 (.926)	−1.208 (1.020)	−1.044 (.898)	−1.426 (.951)
Labor Hours per Establishment (millions)	175 (.285)	−.256 (.237)	.193 (.292)	−.130 (.239)	.036 (.167)	−.153 (.145)	.063 (.151)	−.047 (.133)
Mean Capital-Labor-Hours Ratio (Historical or 1972 $)	026 (.022)	.023 (.025)	.024 (022)	.020 (.025)	.008 (.010)	.019 (.010)	.005 (.010)	.011 (.009)
R^2 (unadjusted)	.553	–	.554	–	.751	–	.803	–
Standard Error of Estimate[d]	459	480	462	.494	.287	.313	.260	.288

Note: Standard errors are enclosed in parentheses.

[a]Unweighted mean [*S.D.*] of rehires rate/new hires rate: State 2-Digit SIC = .278 [.171]; 3-Digit SIC = .502 [320].

[b]Unweighted mean [*S.D.*] of dependent variable: State 2-Digit SIC = −1.464 [.622]; 3-Digit SIC = −.822 [.480].

[c]The weighting procedure used in the *WLS* regressions is described in Section IIE.

[d]For the *WLS* regressions, see fn. d in Table 1.

difference between s in union (u) and nonunion (n) establishments is

$$(s_u - s_n) = \alpha(l_u - l_n) - (\tau_u - \tau_n) \quad (3)$$

If $(s_u - s_n) > 0$, the union-nonunion differential in yearly *UI* benefits per employee is larger than the union-nonunion differential in yearly *UI* taxes per employee.

The average $(s_u - s_n)$ in 1968, 1970, and 1972 for comparable union and nonunion manufacturing firms can be approximated with existing data. The benefit α is equal to

the average weekly *UI* benefit paid to a laid-off individual times the number of weeks he or she received *UI* payments per completed spell on layoff. The average weekly *UI* benefit (in 1972 *CPI*-inflated dollars) received in 1968, 1970, and 1972 (by recipients from all industries) was \$54.[31] A very rough estimate of the mean number of weeks that a laid-off manufacturing employee received *UI* benefits per spell on layoff in these years is 3.9.[32] Therefore, a crude estimate of α is \$211.

Estimates of the average monthly union-nonunion layoff rate differential were based on regressions of the same basic form as regression 1 in Tables 1 and 2, which included only the independent variables that had counterparts in the file used to estimate $(\tau_u - \tau_n)$.[33] Under the not completely innocuous assumption that multiplying these differentials by 12 yields differentials in average annual rates, estimates of $(l_u - l_n)$ were derived that equal .223 and .133.

The union-nonunion differential in annual *UI* taxes per employee $(\tau_u - \tau_n)$ was estimated with information on 4,008 manufacturing establishments from the 1968, 1970, and 1972 *EEC* surveys. This was done by fitting equations under which

$$ln(\tau) = f(w, u, \mathbf{x}) \tag{4}$$

where τ is the annual amount of state *UI* taxes per employee paid by an establishment, w is the average hourly wage in the establishment (included to capture earnings-related differences in per laid-off employee *UI* benefits), u is a dummy variable equal to 1 if a majority of the establishment's employees are covered by collective bargaining agreements, and $\mathbf{x}$ is a vector whose elements are: 2- or 3-digit SIC industry, region, and survey year dummy variables and variables giving the total number of hours worked in the establishment, hours worked per employee, and the fraction of the establishment's employees who are nonoffice.

As would be expected if unionized establishments make greater use of layoffs than do comparable nonunion establishments and if *UI* experience rating is not completely imperfect (which it is not), *UI* taxes per employee were somewhat higher in the presence of collective bargaining. With 141 3-digit SIC dummies, the estimated coefficient (standard error) of the coverage by collective bargaining variable was .098 (.033); with 21 2-digit SIC dummies, the estimated coefficient (standard error) was .070 (.031).[34] To obtain comparability with the estimates of $(l_u - l_n)$, the $ln(\tau)$ regression with the 2-digit SIC industry controls was used to provide an estimate of $(\tau_u - \tau_n)$; it equals \$4.37 (in 1972 *CPI*-inflated dollars).

Using the values discussed above in equation (3) implies that the average annual union-nonunion differential in per employee (not per laid-off employee) *UI* subsidy in 1968, 1970, and 1972 for comparable manufacturing firms was between \$24 and \$43. These amounts represent about .4 to .6

[31] The 1968 and 1970 weekly *UI* benefit data came from Manpower Administration (1971, pp 131, 139); the 1972 information came from Manpower Administration (1973, p.4).

[32] In 1968, 1970, and 1972 the average duration of unemployment as of the relevant *CPS* surveys was 9.8 weeks. In 1973 and 1974 the comparable figure was 9 9 weeks. Therefore, it was assumed that the average duration of unemployment for manufacturing workers on layoff would be roughly the same in both periods. Under this assumption, the 1973 and 1974 May *CPS* surveys, which were readily available, were used to approximate an average duration estimate for the earlier years. Relying on estimates (for all those unemployed) presented in Hyman Kaitz, p. 12, the *CPS* figure was multiplied by .49 to get a crude estimate of a completed spell on layoff for manufacturing workers equal to 4.6 weeks. (Related research by David Lilien suggests that this figure is a lower bound.) This approximation was reduced by .7 of a week, a likely upper bound to the amount of time during which no *UI* benefits were received as a result of a "waiting week" in the first spell of unemployment in a given year. The .7 estimate was based on the fact that in 1968, 1970, and 1972, a lower bound to the number of spells of unemployment in a given year for manufacturing workers who were unemployed at some point during the year was 1.42 (see *BLS*, 1970, 1972b, 1974).

[33] To accomplish this, the capital-labor-hours ratio variable was excluded from model 1 in Table 1 and this variable and the fraction male workers variable were excluded from model 1 in Table 2.

[34] The records of a small number of very large establishments were excluded by *BLS* from the *EEC* files because of confidentiality. There seems to be no obvious reason why these exclusions would bias the estimated collective bargaining coverage coefficients by a significant amount.

percent of the average annual after (payroll and income) tax earnings of fully employed manufacturing production workers with three dependents, during the period under analysis.[35] While rewards in this range do not appear large by themselves, they do not appear trivial if compared to crude estimates of the annual per employee loss in wages associated with the union-nonunion differential in the use of layoffs, which range from 1.2 to 2.0 percent of after-tax earning.[36]

IV. Why Unions Choose Layoffs: Senior vs. Junior Workers

One potential reason why layoffs are preferred to reductions in (the growth of) real wage rates or average hours worked in the presence of a union might have to do with the nature of worker-management communication under collective bargaining. It seems plausible that in unionized firms an average worker preference is transmitted to management at the bargaining table, whereas in nonunion firms employee preferences are more likely to be inferred from the actions of marginal workers (quits and new hires).[37] As Freeman (1976a) points out in an important application of Albert Hirschman's "exit-voice" model to the labor market, it is much more likely that the preferences of senior employees will dominate decision making under an average calculus, since senior workers probably are inframarginal. It follows that, if senior and junior workers have different attitudes toward layoffs and alternative adjustment mechanisms, a different adjustment policy can be expected in the presence of collective bargaining. The one chosen under unionism is likely to be more favorable to senior employees.

It is highly probable that senior union members will have a relatively strong preference for a layoff-intensive adjustment policy given that union-management agreements almost always contain layoff provisions which are more favorable to senior than to junior workers. Most importantly, with additional service comes the right to remain employed until employees with less service have been laid off. A subsample (drawn by *BLS*) of the major contracts in effect in 1970–71 mentioning layoff procedures (81 percent of the major contract work force was covered by contracts referring to these procedures) indicates that seniority was the "sole" or "primary" factor in determining layoff rights for 58 percent of the relevant workers; the comparable figure is 78 percent when contracts in which the issue was "subject to local negotiation" are excluded.[38]

While senior members would most likely prefer not to be laid off in many settings, they might welcome a short period on layoff if recall is a near certainty and seniority is retained (which is usually the case),[39] and if regular and supplemental unemployment benefits are obtainable. In addition, they may prefer a short-term layoff to placement in relatively undesirable or low paying jobs. For these reasons, in almost a fourth of the subsample of 1970–71 major contracts with layoff provisions, senior employees were allowed to waive their seniority rights and be laid off, while retaining eligibility for recall (see *BLS*, 1972a, p. 44).

Thus, senior union members typically incur very small or no costs under an adjustment policy based on layoffs. They would usually incur more if the policy were based on an across-the-board reduction in (the growth of) real wage rates or the hours worked by those employed. As has been seen, unions in manufacturing tend to negotiate contracts that hinder the use of these alternatives to layoffs.

[35]Figures on average weekly earnings after federal Social Security and income taxes for 1968, 1970, and 1972, found in *BLS* (1975b, p. 258), were inflated with the *CPI* and multiplied by 52 to obtain the estimate of average annual after tax earnings in 1972 dollars used in the calculations.

[36]These percentages are based on the estimates of $(l_u - l_n)$ presented in this section and the estimate of the average length of completed spells on layoff discussed in fn 32.

[37]For discussion of this point, see Freeman (1976a) and W. Kip Viscusi.

[38]The figures concerning layoff procedures are from *BLS* (1972a, pp. 53–54)

[39]Data on provisions concerning the retention of seniority by those on layoff are found in *BLS* (1972a, p.56).

Hence, it appears that under collective bargaining the choice is made to impose a disproportionately large share of the labor adjustment burden on junior workers through layoffs and a set of pro-senior-worker layoff provisions.

V. Conclusions

Labor adjustment in *U.S.* manufacturing takes a substantially different form under unionism than in nonunion settings. In firms covered by collective bargaining, layoffs are a much more important adjustment mechanism relative to quits, cuts in (the growth of) real wage rates, reductions in average hours, and discharges, than in comparable firms that are nonunion.

Most laid-off union members return to their previous employer after a short spell of unemployment. Hence, during a period of decreased product demand, management in unionized firms stores unneeded labor outside the firm until an upturn dictates its recall. While nonunion firms also use the layoff-recall method of avoiding excess labor capacity to some extent, the outflows and inflows are substantially smaller than under trade unions. In nonunion firms, employment adjustments are more likely to take the form of quits and new hires.

Layoff-intensive adjustment policies are rewarded under the operation of the unemployment insurance system. Because of imperfect experience rating, firms covered by collective bargaining receive a higher per employee *UI* subsidy for their greater use of layoffs. In addition, under the provisions found in most contracts, adjustments through layoffs are much more favorable to senior union members than are adjustments through across-the-board reductions in (the growth of) real wage rates or hours worked. Thus, the choice of layoffs in unionized firms appears to reflect a decision-making process under which the interests of senior inframarginal workers count a great deal.

In short, to understand layoffs and the alternatives in *U.S.* manufacturing, one must first understand trade unions.

REFERENCES

T. Amemiya and F. Nold, "A Modified Logit Model," IMSSS, Stanford Univ. 1973.

M. N. Baily, "On the Theory of Layoffs and Unemployment," *Econometrica*, July 1977, *45*, 1043–84.

Joseph M. Becker, *Guaranteed Income for the Unemployed*, Baltimore 1968.

———, *Experience Rating in Unemployment Insurance*, Baltimore 1972.

J. Berkson, "Application of the Logistic Function to Bio-Assay," *J. Amer. Statist. Assn.*, Sept. 1944, *39*, 357–65.

———, "A Statistically Precise and Relatively Simple Method of Estimating the Bio-Assay with Quantal Response, Based on the Logistic Function," *J. Amer. Statist. Assn.*, Sept. 1953, *48*, 565–99.

———, "Maximum Likelihood and Minimum χ^2 Estimates of the Logistic Function," *J. Amer. Statist. Assn.*, Mar. 1955, *50*, 130–61.

C. Brown and J. Medoff, "Trade Unions in the Production Process," *J. Polit. Econ.*, June 1978, *86*, 355–78.

H. M. Douty, "Cost-of-Living Escalator Clauses and Inflation," prepared for the Executive Office of the President, Council on Wage and Price Stability, Washington 1975.

M. S. Feldstein, "Unemployment Compensation: Adverse Incentives and Distributional Anomalies," *Nat. Tax J.*, June 1974, *27*, 231–44.

———, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers*, Washington 1975, *3*, 725–45.

———, "Temporary Layoffs in the Theory of Unemployment," *J. Polit. Econ.*, Oct. 1976, *84*, 937–57.

R. B. Freeman, (1976a) "Individual Mobility and Union Voice in the Labor Market," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 361–68.

———, (1976b) "The Effect of Unions on Fringe Benefits," unpublished paper, Harvard Inst. Econ. Res. 1976.

———, "The Exit-Voice Tradeoff in the Labor Market: Unionism, Job Tenure, Quits, and Separations," unpublished paper, Harvard Inst. Econ. Res. 1978.

R.B. Freeman and J.L. Medoff, "New Estimates of the Private Sector Unionism in the United States," *Ind. Labor Rel. Rev.*, Jan. 1979, *32*, 143–74.

——— **and** ———, *What Do Unions Do?*, forthcoming.

Albert O. Hirschman, *Exist, Voice, and Loyalty*, Cambridge 1973.

E. L. Hodgens, "Key Changes in Major Pension Plans," *Mon. Labor Rev.*, July 1975, *98*, 22–7.

H. B. Kaitz, "Analyzing the Length of Spells of Unemployment," *Mon. Labor Rev.*, Nov. 1970, *93*, 11–20.

H. Gregg Lewis, *Unionism and Relative Wages in the United States*, Chicago 1963.

D. M. Lilien, "The Cyclical Pattern of Temporary Layoffs in United States Manufacturing," unpublished paper, Univ. California-La Jolla 1977.

Sumner H. Slichter, James J. Healy, and E. Robert Livernash, *The Impact of Collective Bargaining on Management*, Washington 1960.

H. Theil, "On the Estimation of Relationships Involving Qualitative Variables," *Amer. J. Soc.*, July 1970, *76*, 103–54.

W. K. Viscusi, "Unions, Labor Market Structure, and the Welfare Implications of Quality of Work," unpublished paper, Northwestern Univ. 1977.

J. Zalusky, "Shorter Work Years—Early Retirement," *Amer. Federationist*, Aug. 1977, *84*, 4–8.

Jack Faucett Associates, Inc., *Development of Capital Stock Series by Industry Sector*, Washington 1973.

Bureau of Labor Statistics, *Analysis of Layoff, Recall, and Work-Sharing Procedures in Union Contracts*, Bull. 1209, Washington, Mar. 1957.

———, *Supplemental Unemployment Benefit Plans and Wage-Employment Guarantees*, Bull. 1425–3, Washington, June 1965.

———, *Work Experience of the Population in 1968*, Special Labor Force Rept. 115, Washington 1970.

———, (1972a) *Layoff, Recall, and Work-sharing Procedures*, Bull. 1425–13, Washington 1972.

———, (1972b) *Work Experience of the Population in 1970*, Special Labor Force Rept. 141, Washington 1972.

———, *Work Experience of the Population in March 1972*, Special Labor Force Rept. 162, Washington 1974.

———, (1975a) *Characteristics of Major Collective Bargaining Agreements, July 1, 1974*, Bull. 1888, Washington 1975.

———, (1975b) *Handbook of Labor Statistics 1974*, Washington 1975.

———, (1976a) *Employment and Earnings: United States, 1909–75*, Bull. 1312–9, Washington 1976.

———, (1976b) *BLS Handbook of Methods*, Bull. 1910, Washington, 1976.

———, *Employment and Earnings*, *24*, Sept. 1977, Washington.

Manpower Administration, *Handbook of Unemployment Insurance Financial Data, 1938–1970*, Washington 1971.

———, "Unemployment Insurance Program Letter No. 1251," Washington, Nov. 14, 1973.

U.S. Bureau of the Census, *Current Population Survey*, May 1973; 1974; 1975 (data tape).

———, *Annual Survey of Manufactures*, Washington, various years.

———, *1967 Census of Manufactures*, Washington 1971.

U.S. Department of Commerce, *Industry Profiles 1958–1969*, Washington 1971.

Testing the Theory of Social Security and Life Cycle Accumulation

By LAURENCE J. KOTLIKOFF*

Since its inception the Social Security system has engaged in resource transfers of three kinds, intergenerational, intragenerational, and intertemporal. Transfer of resources across generations, the consumption loan feature of the system, began in 1939 with the payment of benefits to elderly citizens who had paid little or nothing into the system. The 1939 and subsequent amendments to the Social Security Act also weakened the link between taxes paid and benefits received within generations. Within a generation, dependent and surviving widow and widower benefits lead to resource transfers from single to married households and from two-earner households to single earner households. The third resource transfer, the intertemporal transfer, involves simply a reduction in resources when young due to Social Security taxation and an increase in resources when old, the receipt of Social Security benefits.

The impact of these three transfers on the historic level of aggregate savings and hence the size of the current capital stock has been the subject of much recent debate. The unfunded financing of the social insurance program is central to the discussion. While the original 1935 legislation authorized the accumulation of a large trust fund, this goal was essentially abandoned with the 1939 amendments. The failure to accumulate a trust fund, the pay-as-you-go feature of the system is, of course, equivalent to the intergenerational resource transfer. It is argued that this transfer is responsible for an historic reduction in savings relative to consumption.

The theoretical impact of these three types of resource transfers is, however, quite model dependent. For example, a simple Keynesian consumption function with a constant and identical marginal propensity to save out of disposable income for all age groups predicts no change in aggregate savings arising from Social Security resource transfers. A life cycle model of accumulation has, on the other hand, quite different implications. Within a simple life cycle model the introduction of an unfunded Social Security system characterized by a 10 percent tax rate reduces the steady-state capital stock by about 20 percent in general equilibrium and 40 percent in partial equilibrium (see the author). One prerequisite to the resolution of Social Security's historic impact on capital accumulation is, therefore, the empirical verification of micro-economic behavioral responses to Social Security.

This paper presents new micro evidence on the accumulation response of households to Social Security. It is organized in the following manner: Section I reviews the theory of Social Security and life cycle savings: considered here will be the one-for-one replacement of accumulated Social Security taxes for accumulated private savings, the retirement effect, and the effect of changes in lifetime wealth due to the yield of the Social Security system. In Section II econometric specification is used to test the theory. Section III discusses the sample selected from the National Longitudinal Survey (*NLS*) of men aged 45–59, and Section IV presents the empirical findings.

I. The Theory of Social Security and Life Cycle Savings[1]

The different effects of Social Security on life cycle accumulation are easily understood

*University of California-Los Angeles. I have benefitted greatly from discussions with Martin Feldstein, Gary Chamberlain, Steven Marglin, Christophe Chamley, Anthony Pellechio, George von Furstenburg, and Peter Diamond. I wish to thank George Borts and an anonymous referee for helpful suggestions.

[1]The pioneering work on this subject is by Martin Feldstein (1974).

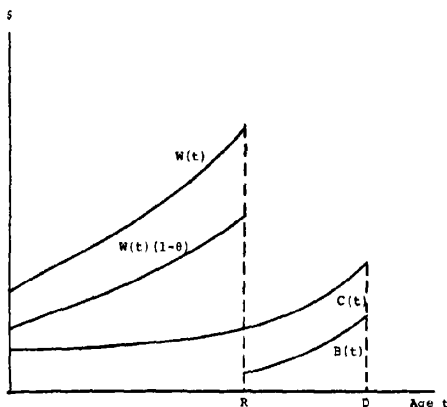



FIGURE 1

with the help of Figure 1. In the diagram a representative life cycler with (for simplicity) a fixed life span D faces, in the absence of Social Security, an earnings stream such as $W(t)$ and chooses a consumption stream such as $C(t)$ and an age of retirement such as R. The choice of consumption at every age t as well as the age of retirement arises from the maximization of an intertemporal utility function of consumption and leisure subject to the following budget constraint (assuming no bequests):

$$(1)\qquad \int_0^D C(t)e^{-rt}\,dt = \int_0^R W(t)e^{-rt}\,dt$$

where r = the rate of interest.

At the micro level the introduction of Social Security reduces the earnings profile by the amount of the Social Security tax $\theta \cdot W(t)$ prior to retirement and provides a Social Security benefit stream $B(t)$ thereafter. The term $\theta \cdot W(t)$ refers to the combined tax on employers and employees. The new budget constraint facing the individual is

$$(2)\qquad \int_0^D C(t)e^{-rt}\,dt = \int_0^{R'} W(t)(1-\theta)e^{-rt}\,dt + \int_{R'}^D B(t)e^{-rt}\,dt$$

If retirement does not change ($R = R'$) and the Social Security system offers an implicit yield on paid taxes equal to the market rate of interest r, then lifetime wealth is not affected by the Social Security system and the consumption profile is unaltered. Under these assumptions accumulated Social Security taxes simply replace accumulated private savings dollar-for-dollar prior to retirement age R.[2] This will be referred to below as the replacement effect.

The assumption of a Social Security yield equal to the rate of interest requires

$$(3)\qquad \int_0^R W(t)\theta e^{-rt}\,dt = \int_R^D B(t)e^{-rt}\,dt$$

Equation (3) states that the present value of lifetime taxes paid in must equal the present value of lifetime benefits received for the system to be actuarially fair. The failure of this equation to hold implies either lifetime wealth increments or decrements due to the Social Security system. Certainly the vast majority of Social Security beneficiaries up to the present have enjoyed positive lifetime wealth increments from the system due on the one hand to high real benefit levels and, on the other hand, to their escape from taxation when young; so far no cohort has paid Social Security taxes for more than forty years. The life cycle model predicts an upward shift in the consumption profile in responses to a positive lifetime wealth increment, assuming consumption at every age is a normal good. This lifetime wealth increment would reduce savings at every age implying a greater than one for one reduction in private accumulation. The magnitude of this departure from the

[2]To see this let A_T be the value of private assets at time T, then

$$A_T + \int_R^D B(t)e^{-r(t-T)}\,dt + \int_T^R W(t)(1-\theta)e^{-r(t-T)}\,dt = \int_T^D C(t)e^{-r(t-T)}\,dt$$

Future consumption is financed out of current private assets, future Social Security benefits, and net future earnings. Using equation (3) we may rewrite the expression as

$$A_T + \int_0^T \theta W(t)e^{r(T-t)}\,dt = \int_T^D C(t)e^{-r(t-T)}\,dt - \int_T^R W(t)e^{-r(t-T)}\,dt$$

Since the right-hand side is independent of Social Security, private assets are offset dollar-for-dollar by accumulated Social Security taxes.

one-for-one replacement of private net worth by accumulated Social Security taxes is limited at any age to some fraction of the lifetime wealth increment evaluated at age zero (LWI_0). For example, if the increase in consumption financed by the LWI_0 was a constant ΔC at every age, then

$$\text{(4)} \qquad \int_0^D \Delta C e^{-rt}\, dt = LWI_0$$

or

$$\Delta C = \frac{r}{1 - e^{-rD}} LWI_0$$

In terms of the diagram, the consumption profile would shift upward by an amount $rLWI_0/(1 - e^{-rD})$ at every age. By age x the fraction of LWI_0 consumed equals $(1 - e^{-rx})/(1 - e^{-rD})$. The effect of this additional consumption on wealth holdings at age x is found by accumulating forward up to age x the reduction in savings due to the LWI_0 and may be expressed as

$$\text{(5)} \qquad \int_0^x \Delta S e^{rt}\, dt = \frac{-(1 - e^{-rx})}{(1 - e^{-rD})} e^{rx} LWI_0 = \frac{-(1 - e^{-rx})}{(1 - e^{-rD})} LWI_x$$

LWI_x is the value at age x of the lifetime wealth increment receivable at age zero. Economic theory suggests then that the inclusion of the variable LWI_x in an accumulation regression should yield a coefficient ranging between 0 and -1 depending on the age of the respondent. For example, in the case of constant incremental consumption at every age x, the term $(1 - e^{-rx})/(1 - e^{-rD})$ takes the value .68 for an age of death of 55, x equal to 30, and an interest rate of .02. Since age zero in the life cycle model corresponds not to birth but rather to the beginning of one's productive life, say age 20, .68 is the fraction of LWI_x a 50-year old would have consumed who expected to live until age 75.[3]

[3]The presence of inflation does not alter any of these conclusions since all the variables in the preceding equations may be taken as real including the rate of interest. In this empirical work all variables are measured in real 1966 dollars.

Both the intergenerational and intragenerational resource transfers are captured by the LWI_x variable. The intertemporal transfer, holding LWI_x constant, is neatly summarized by the amount of accumulated Social Security taxes ($ASST$). Again the intertemporal transfer refers to the payment of taxes when young and the transfer back of these taxes when old in the form of benefits with an implied return equal to the market rate of interest. *Ceteris paribus*, the predicted coefficient of $ASST$ in an accumulation regression should equal -1. The life cycle model predicts explicit signs and magnitudes of coefficients for resource transfer variables within an annual savings regression as well. In an annual regression of savings every dollar of Social Security taxes should reduce annual private savings by more than one dollar holding LWI_x fixed, since disposable income is reduced by the amount of interest on accumulated Social Security tax payments. Holding LWI_x fixed, the annual reduction in savings for the preretirement age group should equal the annual amount of taxes plus the interest on accumulated Social Security taxes. A dollar increase in LWI_x should reduce annual savings by a fraction of a dollar depending on the age of the respondent.

Thus far I have not considered the impact of Social Security on the age of retirement and through retirement age on accumulation. As Martin Feldstein has pointed out, Social Security may induce early retirement due to an implicit tax on post-62 earnings;[4] the savings response of the young to a planned earlier retirement is likely to be positive. A reduction in the age of retirement will reduce lifetime wealth by shortening the earnings stream. Under reasonable assumptions this will shift the consumption profile downwards increasing savings at young ages. The magni-

[4]The "actuarial" reduction of benefits for those choosing early retirement may imply a zero implicit tax between the ages 62–65 to the extent that the actuarial reduction is truly fair. To be truly fair benefits levels must rise to compensate for the postponement, the higher risks of mortality, as well as the additional tax contributions paid in during these years. For a comprehensive review of the growing literature on Social Security and retirement see Colin Cambell and Rosemary Cambell.

tude of the retirement effect is critical in determining Social Security's net impact on accumulation. The argument for an historical reduction in aggregate *U.S.* savings relative to consumption rests on an increase in consumption due to intergenerational transfers (the lifetime wealth increment effect) by the initial older generation without a completely offsetting reduction in consumption because of earlier anticipated retirement by the young. Note that the intertemporal transfer, the replacement effect, implies by itself no change in aggregate consumption according to the life cycle model since it involves no change in anyone's lifetime budget constraint.[5]

The above discussion concerning the lifetime wealth increment and replacement characteristics of Social Security is subject to qualification in the case of imperfect capital markets. If capital markets are imperfect and one cannot borrow against future Social Security benefits, fixed savings goals (for example, a downpayment on a house) prior to retirement may lead to a one-for-one reduction in consumption up to a certain age and an increase in consumption compared to previous levels thereafter. In other words the consumption profile could rotate counter clockwise at a given age. Variables to test for this capital market imperfection will be suggested below.

II. Econometric Specification

To test empirically the micro response to Social Security I specify two linear regressions, one for household wealth accumulation prior to age of retirement and one for the expected age of retirement. The basic framework for the accumulation equation is given by

$$(6) \quad Net\ Worth = B_1 + B_2 ASST + B_3 LWI_x + B_4 RETAGE + B_6 LTLABI + \gamma' Z + E$$

where *Net Worth* = assets less liabilities of the household.

ASST = the value to the present of household accumulated Social Security taxes; that is, paid in employee plus employer Social Security taxes are accumulated up to the present at the market rate of interest. Households here are taken to be the husband and wife if married, otherwise the single household head. Social Security taxes and benefits of other family members are not considered.

LWI_x = the absolute dollar yield of the Social Security system to the individual household in current dollars and is equal to the present expected value of future Social Security benefits less the present expected value of future Social Security taxes less the value of past paid in Social Security taxes accumulated up to the present. Letting *PVFB* and *PVFT* stand for the present value of future benefits and taxes, respectively,

$$LWI_x = PVFB - PVFT - ASST$$

If LWI_x is positive, the household has received a higher yield on its taxes from the Social Security system than the market rate of return.

RETAGE = the expected age of retirement of head.

LTLABI = the current dollar value of household lifetime gross labor income; where gross refers to gross of employer plus employee Social Security taxes although net of income taxes.

$\gamma' Z$ = the vector sum of additional exogenous variables.

E = the error of the regression.

Recall that the expected coefficient on the term *ASST* is -1: to see this most clearly, compare two families. Family *A* is insured by Social Security and family *B* is not insured. Both families have identical streams of gross labor income. Assume all other characteristics of the two families are identical and that the family with Social Security has an LWI_x equal to zero. Assuming either no capital

[5]The intertemporal transfer would occur whether or not the system was funded. For example, under a fully funded system we would still observe people paying in taxes when young and drawing out benefits when old. Under a fully funded system, however, the benefits drawn out would equal one's own contribution plus interest, while under an unfunded system the benefits drawn out would correspond to the tax contribution of the succeeding generation.

market constraints to borrowing against future Social Security benefits or that such borrowing is not desired, the consumption streams of the two families are identical since they are both based on the same lifetime wealth. Hence, the sum of accumulated savings in both private (*Net Worth*) and Social Security forms (*ASST*) will be identical for both families and the coefficient on *ASST* should equal -1. (The predicted partial coefficient for *ASST* is -1 regardless of the value of LWI_x.)[6] Let us now make the assumption that family *A* has a positive value of LWI_x. Family *A* has therefore a higher lifetime wealth than does family *B*. Family *A* will consume more at any point in time than family *B* assuming noninferiority of consumption at any point in time. Prior to retirement family *A* will have less private plus Social Security accumulation (*Net Worth* plus *ASST*) than family *B* since they have consumed more each year for identical gross labor incomes. The coefficient in LWI_x should therefore be negative. In addition this coefficient should be less than 1 in absolute value since only a portion of this lifetime wealth increment will have been consumed by family *A* each year up to the current period. A significant negative coefficient on the lifetime wealth increment variable is critical to the argument that Social Security has reduced aggregate savings. A negative coefficient on *ASST* for our preretirement sample would not by itself imply a reduction in aggregate savings since even a simple Keynesian savings function of the form $S = a + bY_d$ (where S is savings, Y_d is disposable income, and a, b are coefficients) would predict a reduction in preretirement savings because of the reduction in disposable income from the Social Security tax.

Ideally equation (6) should include the expected retirement age of the spouse as well as that of the head, *RETAGE*. Unfortunately the *NLS* data reports this information only for the head. In addition the data does not report information about earnings histories. Hence in our regression we use a proxy for *LTLABI* a two-year average of disposable labor income adding in the employer's tax contribution to Social Security based on the two-year earnings average and deducting an estimate of income taxes paid on that labor income.[7] A recent study by James Adams points to an elasticity of lifetime bequests with respect to lifetime wealth in excess of unity. To capture rising accumulation for bequests as lifetime resources increase the square of this constructed average gross labor income (henceforth *ALY* and *ALY2*) is included in the regression. Since this proxy presumably measures *LTLABI* with error the education of the head is included as an explanatory variable. An additional bequest variable *IHER* was coded 1 for respondents indicating a desire to leave an inheritance.

The savings behavior of two-earner families may differ from that of single earner families for at least two reasons. First two-earner families have greater work-related expenses; hence for a given level of *ALY* the disposable labor income of two-earner families is less than that of single earner households. Second the riskiness of the future labor earnings stream from death, disability, or loss of work is smaller when the labor income is divided among two earners rather than one. Hence two-earner families will have a smaller precautionary motive for savings. To allow for the possibility that two-earner families save less we include the variables *ALYT* and *ALY2T* which equal *ALY* and *ALY2* multiplied by a dummy variable for two-earner households.

Other variables included in the regression are dummies for marriage (*MAR*), race (*RACE*), and heads who are separated, widowed, or divorced (*SWD*). In addition the head and wife's ages and the square of their ages (*AGEH*, *AGEH2*, *AGEW*, *AGEW2*) as well as the number of family members (*SIZE*) are exogenous variables. Ideally one would like to treat pension contributions in an identical manner as Social Security taxes, accumulating them up at the market rate of

[6]To see this alter (3) to let the lifetime wealth increment equal the difference between present benefits and present taxes and proceed as in fn. 2.

[7]Presumably, reported labor income is gross of the employee's Social Security tax contribution but net of the employer's; hence, we need add in only the employer's half to obtain gross labor income.

interest and testing for a coefficient of minus one. Unfortunately the data do not provide reliable information for current, let alone past, pension contributions. Hence two dummy variables are introduced to capture the average effects of pensions on the accumulation of wealth: *PEN* takes the value 1 if the respondent reported participation in a pension plan; since government workers have more generous and secure pensions, *GPEN* was coded 1 if in addition the respondent was a government worker.

So far I have presented variables to capture three of the four theoretical points in Section I, viz., replacement, lifetime wealth increment, and retirement. The fourth issue pertains to capital market constraints on borrowing against future Social Security benefits. Families with fixed accumulation goals such as a downpayment for a house may be forced to reduce consumption in response to the Social Security tax if these taxes constitute a large fraction of their pre-Social Security savings. For such families the coefficient of the Social Security variables *ASST* and LWI_x may be zero. To test for this possibility two additional variables are used in separate regressions below. They are *HASST* and $HLTW_x$, dummy variables for home ownership multiplied by *ASST* and LTW_x. The coefficients of *HASST* and *ASST* should be positive and negative one, respectively. The coefficients on $HLTW_x$ and LTW_x should also be of opposite sign, equal, and less than one in absolute value. Hence, when homeownership is indicated the two sets of variables will cancel implying zero effect of Social Security on accumulation.[8]

A. *Social Security and Retirement Intentions*

The *NLS* data provide an excellent opportunity to test whether the Social Security system does indeed reduce the intended age of retirement. The decision to retire rests on a comparison of the shadow price of leisure at a given age with the wage. A number of factors enter the calculations of the shadow price of leisure including health, work attitudes, marital status, and age. Certainly a key to the decision is the level of Social Security benefits available when working compared to the available benefits when retired. For workers between the ages 65 and 72 the Social Security earnings test represents an implicit tax on labor supply (see Michael Boskin). Currently, benefits for this age group are reduced by 50 cents for every dollar earned over $4,000. Beyond the age of 72 all workers are entitled to full benefits independent of retirement. Between the ages 62 and 65 the same earnings test applies; however, available benefits are "actuarially" reduced 5/9 of 1 percent for each month that benefits are received before age 65. Thus a worker retiring at age 62 receives a benefit which is permanently 20 percent lower than the benefit he would receive if he first retires at age 65. The "actuarial" reduction, if it is truly actuarially fair, implies that the Social Security earnings test does not represent an implicit tax on work effort between 62 and 65; foregoing benefits at these ages will result in higher benefits in later years when actual retirement occurs, leaving the present expected value of benefits as of age 62 the same independent of age of retirement between 62 and 65. Given the extent of early retirement between 62 and 65, the possibility that workers either do not have knowledge of or do not understand actuarial reduction must be strongly entertained.[9] Our econometric analysis will consider each possibility, perception of actuarial reduction and nonperception of actuarial reduction in turn.

The non-linear Social Security earnings test schedule implies that the implicit marginal tax on an additional hour of work depends itself on the extent of earnings as well as the age. If our data on expected future work effort were sufficiently rich, a maximum likelihood technique could be employed assessing the probability of a worker's locating on a given branch of his budget frontier at a given age. Given the data limitations the

[8]This assumes that savings for fixed savings goals occurs up to the current age of the respondent.

[9]Another explanation for Social Security induced retirement between 62 and 65 is the inability to borrow against Social Security benefits. Note, there is a trivial 1 percent annual reduction in benefits for retirement prior to age 72 but after age 65.

econometric path chosen was to specify a Social Security tax variable (*SBENL*) defined as the ratio of Social Security benefits lost at full-time work to full-time earnings. While this tax variable obviously does not capture the full complexity of the kinky budget frontier, it distinquishes quite well workers facing high implicit taxes in the neighborhood of full-time work. Since the dependent variable to be explained is the expected age of retirement, that is, the expected age at which full-time work will cease, this tax variable seems quite appropriate. The basic framework for the expected retirement age regression can now be written as

$$(7) \quad RETAGE = \alpha_1 + \alpha_2 SBENL + \alpha_3 LWI_x + \alpha_4 LTLABI + \delta' H + u$$

where $\delta' H$ is a vector of additional exogenous variables and u is the regression error. The specification of (7) is suggested by the life cycle theory which relates the endogenous decisions of accumulation and work effort to the exogenous variables of lifetime wealth and provisions of the Social Security system. Indeed since the error term in (7) may be correlated with E in (6) we shall consider a two-stage estimation of (6) as well as simple *OLS*. Even under the assumption of nonperception of actuarial reduction, (7) is inappropriately specified for the entire sample. Prior to age 62 there is no implicit Social Security tax on work effort. Hence, *OLS* estimation of (7) over the entire sample will yield an estimate of the tax rate coefficient α_2 biased toward zero; given the decision to retire prior to age 62, the choice of the exact age to retire before 62 is independent of the tax rate one would face after age 62. In the case that actuarial reduction is perceived, the decision to retire prior to age 65 is independent of the tax rate one would face after age 65. To test this possible bias on α_2 equation (7) was run for samples with expected retirement ages greater than 62 and 65 correcting for the sample selection bias introduced from truncating the sample on the dependent variable.[10]

The exogenous variables of the expected retirement age regression not yet identified include dummies for race, marriage, and separated, widowed, or divorced heads; age and education of the head; two pension dummies mentioned above; and number of family members. In addition there are three health dummies *HG*, *HF*, and *HP*, corresponding to the respondent's assessment of his health as good, fair, or poor, the excluded category being excellent health. Finally, three dummy variables are included to capture work attitudes and attachment: *PROF* takes the value 1 for respondents reporting professional or managerial occupations; *ATDJ* corresponds to attitude toward current job and is coded 1 for people who indicated disliking their job either somewhat or very much; *ATDW* takes the value 1 for respondents answering no to the question, "If you could live comfortably without working would you still work?"

III. The National Longitudinal Sample

Beginning in 1966 the Bureau of Census conducted a series of surveys of male household heads age 45–59. While the surveys deal primarily with labor market questions, a rich

[10]See James Heckman. The procedure is simply to take account of the fact that the error u in (7) has a nonzero mean given the sample selection rule. An additional variable $E(u/\text{sample selection})$ is added to the list of exogenous variables (written compactly here as $\lambda' Z$) to form the regression: $RETAGE = \lambda' Z + E(u/\text{sample selection}) + v$. Since $v = u - E(u/\text{sample selection})$ the expectation of v over the selected sample is zero and the above equation may be estimated by least squares yielding consistent estimates for the coefficients in λ In forming the term $E(u/\text{sample selection})$ let us follow the labor leisure choice literature. It is assumed that the expectation of full-time work beyond the attainment of age 65 is based on a comparison of the shadow price of leisure expected to prevail at age 65, S^*_{65}, with the expected age 65 net wage W^{n}_{65}. The decision to retire implies that the shadow price of leisure at forty hours of work exceeds the wage. This comparison gives rise to a probit regression from which $E(u/\text{sample selection})$, which is called a mills ratio, is estimated up to a constant; that is, the coefficients obtained in the probit regression are used in forming the mills ratio.

amount of information was collected concerning financial status, retirement plans, and eligibility for private pensions and Social Security. The data used in this study come almost exclusively from the 1966 survey. The 1966 survey also asks extensive questions about earnings in 1965. The dependent variable, household accumulated wealth, is one of the key variables recreated by statisticians in the U.S. Department of Labor. This variable falls short of the economist's definition of net worth since it fails to include the cash value of equity in life insurance and the value of consumer durables.

Information detailing the expected age of retirement is of two kinds. Either the respondent stated an actual age, or he indicated the intention never to retire (this group represents 14 percent of respondents in our sample). In the accumulation regression the *RETAGE* variable is replaced by *RET*, the actual retirement age when indicated or by *NORET*, a dummy taking the value 1 when the respondent stated he would never retire. For purpose of the expected retirement age regressions, an expected age of retirement of 70 was assigned to this second group.

Since the construction of the Social Security variables (see the Appendix) is based in large part on estimates of labor income, only those observations reporting positive labor income for working heads in both 1965 and 1966 were included; self-employed heads were excluded since their reported wage income may include a return to capital. After several additional consistency checks the final sample totaled 2,124.

In Table 1, I present the distribution of computed household lifetime wealth increments (LTW_x) by age of head, marital status, and average household labor income (ALY). The table is based on 2,587 observations; both households with covered and uncovered heads are included; the distribution excluding uncovered heads is quite similar.

The intergenerational transfers in Table 1 are quite large when compared with either the mean value of average household labor income $7,000, or the mean value of household net worth, $15,000. Perhaps the most striking feature of Table 1 is the unequal treatment of married and single households. The average married lifetime wealth increment of $10,431 is 3.7 times the average for

TABLE 1—MEAN HOUSEHOLD LIFETIME WEALTH INCREMENTS[a]

Household Average Labor Income	45–50		51–54		55–59		Total	
	M	S	M	S	M	S	M	S
0–3,000	8997	2647	9998	3838	11310	4652	10023	3668
	(1632)	(664)	(2364)	(542)	(3193)	(921)	(2418)	(718)
3–6,000	9880	1046	12048	3049	15588	5542	11910	3032
	(2364)	(852)	(3120)	(904)	(3133)	(1315)	(2783)	(1028)
6–10,000	8599	−273	11441	1748	14720	4968	10821	1671
	(2541)	(796)	(2558)	(965)	(3282)	(1457)	(2725)	(1036)
10–15,000	6705	−458[b]	9578	–	13018	–	8677	1071[b]
	(3358)	(729)	(3180)		(3593)		(3352)	(1179)
15–25,000	6447	–	9068	–	13378	–	8948	–
	(3779)		(3475)		(3375)		(3583)	
25,000+	8036[b]	–	10845[b]	–	13744[b]	–	10385	–
	(3952)		(3732)		(3920)		(3744)	
Total	8330	1058	10954	2929	14098	5102	10431	2802
	(2768)	(763)	(2844)	(779)	(3290)	(1249)	(2919)	(936)

Note: Standard deviations are in parentheses. Dashes indicate fewer than five observations.

[a]*M* and *S* stand for married and single, respectively.

[b]Fewer than fifteen observations.

single heads, \$2,802. If we divide the married figure by two it is clear that for some cells the gain in becoming married is as much as \$4,000 per person.

The lifetime wealth increments increase by age for two reasons. The older the cohorts the fewer the number of years exposed to the Social Security tax; 59 year olds in the sample were age 30 at the initiation of the Social Security system and thus escaped ten years of tax payments. The second reason for the sharp increase in LWI_x with age is that for a given age-zero lifetime wealth increment (LTW_0), the older the cohort the longer the accrued interest on the LTW_0; i.e., $LWI_x = e^{rx} LWI_0$. For example, if a 50-year old and a 40-year old both had a lifetime wealth increment of \$10,000 evaluated in present dollars as of age zero, the value of the \$10,000 evaluated in present dollars as of age 50 must exceed the \$10,000 evaluated in present dollars as of age 40. To make the figures for the older 55–59 age group more comparable to those of the 45–50 we can ask what the value of the former group's increment was ten years earlier when the cohort was 45–49. Discounting \$14,098 at 3 percent for ten years yields a figure of \$10,432, \$2,102 greater than the \$8,330 figure for the 45–50 group. This differential is now solely a function of the number of years of exposure to the Social Security tax. Table 1 indicates that as a fraction of labor income the lifetime wealth increments arising from Social Security are progressive. The standard deviations are substantial and point to sizeable within cell inequalities.

TABLE 2—REGRESSION COEFFICIENTS FROM ACCUMULATION EQUATION

Variable	Mean	Coefficient	Standard Error
ASST	6629	− 666	.305
LWIx	9591	.237	.202
ALY	6994	1.25	.598
ALY2	621 *D*+8	158 *D*−3	250 *D*−4
ALYT	3799	004	.327
ALY2T	351 *D*+8	−.684 *D*−4	.256 *D*−4
AGEH	51 1	−2756	3059.
AGEH2	2633	30.6	29 7
AGEW	42 1	105.3	363
AGEW2	2016.	1 72	4.25
PEN	647	−1054	1221.
GPEN	147	141	1500
RET	55 0	−428	150
NORET	138	−27455	9714
MAR	894	−7997	8747
EDH	9 46	489	163
RACE	686	3482	1190
SWD	075	−2801	3263
IHER	595	1325	983
SIZE	4 67	−370	−1 71
Constant	1 00	82609	79027

$R^2 =$ 330
$F(20,2104) =$ 51 687
Mean of Dependent Variable = 15098

Note: Variables are defined in the text The notation $D + a$ means the coefficient is multiplied by 10^a.

IV. The Empirical Findings

Table 2 reports the regression results for the accumulation regression specified above as equation (6). The regression is highly significant with an R^2 of .332. The coefficient of accumulated Social Security taxes is −.666 with a standard error of .305. This coefficient is significantly different from zero and lies within 1.3 standard deviations of −1, the prediction of the life cycle theory. The coefficient for LWI_x, .237, on the other hand, differs significantly from a predicted negative fraction of about −.68. One interpretation of this insignificant coefficient for LWI_x is simply that households fail to accurately foresee their future benefits prior to age of retirement. Accurate projection of Social Security benefits requires detailed knowledge of the dependent and surviving spouse benefit provisions, an assessment of husband and wife survival probabilities at different ages, an understanding of current benefit levels, and some notion of the future growth rate of these benefits. In the absence of such knowledge, households may simply assume they will receive the market yield on their tax contributions, that is, that LWI_x is zero. This, of course, casts doubt on the validity of life cycle model in general, since the life cycle model requires a great deal of foresight if it is to be valid. Objections may be raised to including uncovered respondents in the regression since they account for a large part of the variance in the Social Security variables and are primarily government employees who may become eligible for Social Security in the future. Excluding this group leads to coeffi-

cient values of $-.902$ ($t = 1.977$) for *ASST* and .189 ($t = .065$) for LWI_x.

Turning to the retirement variables, the coefficients for *RET* and *NORET* are each significantly negative as predicted by the life cycle theory. The expectation of a year's earlier retirement increases accumulation by about $428. Multiplying $428 by 63.8, the average expected age of retirement for those stating an explicit expectation, yields $27,306. This figure is $149 less than the coefficient for *NORET*, a dummy for respondents who state they will never retire. These respondents appear, then, to accumulate only $149 less than respondents expecting to retire at age 63.8.[11] The $428 figure appears small when compared to the mean value of the heads average labor income, $6,034. Four factors are pertinent to the evaluation of this figure. First the average head's age in the sample is 51.1 (12.7 years less than the average expected age of retirement for those with retirement expectations). Hence a typical respondent has an additional thirteen years to save for his retirement period.[12] Secondly, Social Security benefits will replace foregone earnings for retirement ages greater than 65 beyond which actuarial reduction does not occur. Third, the stated intention to retire does not necessarily mean the respondent plans to stop working altogether. Forty percent of new male and 34 percent of new female Social Security beneficiaries reported some employment in a 1968 survey. (See Patience Lauriat and William Rabin, p. 8.) Finally, these retirement expectations are formed under considerable uncertainty about future health and family needs. Indeed the correlation between retirement intentions and actual retirement is not high.

The *NLS* data permits a comparison of expected with actual retirement behavior. The expected retirement age stated in 1966 was compared with 1973 retirement expectations and 1973 employment status for 1,787 respondents who appeared in both the 1966 and 1973 surveys. Of the 1,787 observations, 369 (21 percent) exhibited employment behavior at variance with their stated 1966 expectation. An additional 570 observations (32 percent) changed their expected age of retirement by at least one year. The figures for the 56–59 age group are more revealing since a larger percentage of this group had the opportunity to demonstrate employment behavior at odds with their 1966 expectations. For this group 38 percent exhibited inconsistent behavior and another 8.6 percent changed their expectations about retirement. The change in expected retirement age averaged 4.0 years for those who revised upwards their retirement expectation and 3.7 years for those who revised downward (after excluding those expecting never to retire either 1966 or 1973). These figures suggest that the expectations formed in 1966 are best characterized as guesses rather than firm plans. Serious planning and saving for retirement may occur only a few years prior to the actual reduction in work effort. Given the uncertainty with which these retirement expectations appear to be held, the small coefficient on *RET* is not surprising.

The size of the *RET* coefficient rules out the possibility that increased savings due to induced earlier retirement substantially offsets the replacement of private savings by Social Security taxes. Taking the point estimate of the *ASST* coefficient, -0.666, the average reduction in private accumulation from this variable was $4,415. To offset this reduction expected retirement age must fall by an implausible 10.32 = $4,415/$428 years.

The coefficient of the labor income variables *ALY*, and its square *ALY2*, and those of *ALYT* and *ALY2T* (the *ALY* and *ALY2* variables times a dummy for two-earner households) exhibit predicted signs and reasonable magnitudes. At the mean level of *ALY* an additional dollar of average labor income raises accumulation by $3.46 for single earner families and $2.51 for two-

[11]This comparison simply uses the regression equation to determine the difference in predicted net worth for two respondents, identical in all respects, except that one expects to retire at age 63.8 and the other expects never to retire.

[12]This is confirmed by respecifying *RET* as *RET* times three dummies for age 45–50, 51–55, and 56–59. The coefficients obtained are -406, -449, and -466, respectively.

earner families. The point elasticities of accumulation with respect to *ALY* are 1.59 for single earner and 1.15 for two-earner households and accord well with wealth elasticities of bequest estimated in Adams. While the inclusion of *ALYT* and *ALY2T* seems justified on theoretical grounds we caution that the coefficients of *ASST* and LWI_x are highly sensitive to these variables. Omitting these variables leads to coefficients of .686 ($t = 3.88$) for LWI_x and -1.270 ($t = -4.71$) for *ASST*.

The effects of both the head's education and race on accumulation are measured quite precisely. Whether these variables independently influence life cycle accumulation or are simply correlated with the error in the proxy for lifetime earnings *ALY* is impossible to say. The coefficient on household size, -370, ($t = 1.71$) probably reflects offsetting consumption expenditures for children and increased savings for their education.

Surprisingly, neither the pension dummy *PEN* nor its interaction with a dummy for government workers is significantly negative. Allowing for interaction of these variables with *ALY* failed to yield significantly negative coefficients. These findings differ from those of Alicia Munnell (1976) who used the same data set to run an annual savings regression. Rather than examine Munnell's specification in detail I report findings on a savings regression of my own derived from differentiating the accumulation equation with respect to time. The dependent savings variable is defined as in Munnell by the difference in net worth in 1969 and 1966 divided by 3. In the savings equation *ASST* is replaced by *SSTX*, household Social Security tax contributions. The coefficient for *SSTX* is -2.42 but is measured very imprecisely; the standard error is 2.53. The LWI_x coefficient again displays the wrong sign (.203 with a standard of .099). In addition neither of the pension dummies is significantly negative; indeed *GPEN* is significantly positive. The overall explanatory power of the exogeneous variables is quite low, the $R^2 = .046$. The coefficient of LWI_x suggests that my savings equation provides no support for a reduction in aggregate savings due to the introduction of unfunded Social Security.

Returning to the accumulation regression, the capital market constraint variables await discussion. Recall, to test capital market constraints against borrowing we add two variables *HASST* and $HLWI_x$, defined as *ASST* and LWI_x, multiplied by a dummy for home ownership. The results here are quite supportive of the borrowing constraint hypothesis. The coefficients are -1.500 for *ASST* and 1.173 for *HASST* with respective t-values of -4.029 and 3.842. The LWI_x coefficient is .080 ($t = .322$), and the $HLWI_x$ coefficient is .139 ($t = .859$). These findings imply essentially a dollar for dollar replacement of savings by tax contributions for nonhomeowners and a zero reduction in savings for homeowners. Since homeowners constitute a large proportion of this sample (70 percent) and of household heads in general, this finding greatly narrows the scope for an aggregate reduction in savings due to Social Security. However, the potential endogeneity of the homeownership dummy as well as its unproven ability to proxy for households with fixed savings goals facing borrowing constraints cautions against relying too strongly on these results.

One final accumulation regression requires reporting. Since *RET* and *NORET* are potentially endogenous I estimated the accumulation equation with two-stage least squares. In the second stage the coefficient for the predicted retirement age is a positive 2028.5 ($t = 3.342$). The coefficients of *ASST* increased to $-.773$ and that of LWI_x decreased to .093. Other coefficients were essentially unaffected with the exception of the *PEN* variable. This coefficient is now 1915.9 although still insignificant. The R^2 for the first stage retirement age prediction equation is only .131 and presumably accounts for these poor second stage findings.

A. *Expected Retirement Age Regressions*

The coefficients for the expected retirement age regression appear in Table 3. In this regression I ignore the issue of bias in the coefficient of *SBENL* and run over the entire sample. The coefficient of the key Social Security tax rate variable *SBENL*, .626, is insignificant and the wrong sign. Eliminating

TABLE 3—EXPECTED AGE OF RETIREMENT REGRESSIONS

Variable	Coefficient	Standard Error
SBENL	.626	847
LWIx	.570 D−4	.330 D−4
ALY	−.320 D−3	.838 D−4
ALY2	.921 D−8	.374 D−8
ALYT	.733 D−4	.535 D−4
ALY2T	−.238 D−8	.428 D−8
AGEH	.097	.028
AGEW	−.026	.015
PEN	−1.236	.197
GPEN	−.601	.245
SIZE	006	.036
MAR	.092	.896
EDH	−.541 D−2	029
RACE	−.044	198
SWD	−.880	.539
HG	−.036	.175
HF	−.235	252
HP	− 229	.642
ATDJ	−.823	.311
ATDW	−1.595	190
PROF	.758	.251
Constant	62 834	1 575
R^2 = .128		
F(21,2102) = 14.708		
Mean of Dependent Variable = 64 64		

Note Variables defined in the text. The notation $D + a$ means the coefficient is multiplied by 10^a

respondents who reported compulsory retirement ages and those not covered by Social Security produced a negative, although still insignificant −.488 (t = −.475) *SBENL* coefficient. Since the wife can collect dependent benefits only if the husband collects some benefits, *SBENL* was changed to include the wife's dependent benefit less her own benefit for wives no more than five years younger than their husbands. This second lost benefit variable yielded a −.260 coefficient but still insignificant (t = .448). Table 3 indicates that positive lifetime wealth increments captured by LWI_x also do not lead to earlier expected retirement.

In contrast with the Social Security variables the pension dummies are important predictors of expected retirement age. Coverage under a private pension plan entails a 1.2-year earlier expected retirement; for government pension coverage the impact is 1.8 years.

The coefficient on the age of the head is significant and quite interesting. A ten-year increase in age means a 11.6 month later expected date of retirement. Since my comparison of retirement intentions with actual practices indicates that both the young and the old tend to overproject their additional working span, this coefficient probably reflects the general trend among the young to retire earlier.(See *Manpower Report of the President*.) The health and employment attitudinal variables all display the anticipated negative signs, although none of the health dummies is significant. Professionals and managers expect to retire about nine months later than other respondents indicating a lower disutility to work for this group.

B. *Correction for Bias in the Social Security Tax Variable*

The regression in Table 3 is a reduced form even when actuarial reduction is perceived; however, the coefficient on *SBENL* is likely to be biased toward zero as discussed in Section II. Including the term to correct for sample selection bias, and estimating the regression for the sample with expected retirement age greater than 62 failed to yield a significantly negative coefficient for *SBENL*; the estimated coefficient is −1.061 (t = −1.458). Checking for a remaining bias due to actuarial reduction by sampling over the post-65 expected retirement age group also failed to yield a significantly negative coefficient for *SBENL*. For this sample the estimated coefficient is −.069 (t = −.067). At this stage we tentatively conclude that Social Security does not significantly influence the intended age of retirement for our sample of 45–59 year olds. Taking the largest coefficient found for *SBENL*, −1.0612, implies only one-third of a year earlier intended retirement when this tax rate falls from its mean of about .3 to zero. Finally, since one may object to the arbitrary assignment of age 70 to respondents intending never to retire, a simple probit regression was run explaining the choice of intended retirement age—between those who would choose to retire at or before 65, and those who would retire later. Here the variable *SBENL* was introduced independently and yielded a coefficient of −.155 (t = .475).

V. Summary and Conclusion

This paper has examined the theoretical savings response to Social Security implied by the life cycle theory of accumulation. The life cycle theory predicts explicit signs and magnitudes of coefficients for Social Security variables in both accumulated savings and annual savings regressions. In addition, the theory predicts that savings is responsive to retirement plans; these retirement plans may, in turn, be influenced by the availability of Social Security benefits.

The econometric results give mixed support to the notion that the micro-economic mechanisms of the life cycle model are at work. Social Security appears to significantly reduce accumulated savings of this 45–59 year old sample. Whether they view Social Security taxes as equivalent to other taxes or as a significant replacement for private savings remains unclear, however. The point estimate of this replacement (*ASST*) is −.666 which is neither significantly different from the −1 predicted by the life cycle theory or the −.2 to −.3 predicted by a simple Keynesian consumption function. The finding of a positive and insignificant coefficient for LWI_x for this age group runs counter to the life cycle theory. While there is evidence that Social Security tax contributions have reduced private savings of the young, there is no evidence that aggregate saving has been reduced. The poor findings on the lifetime wealth increment imply that the savings of the old may have increased to offset the reduced savings of the young leaving zero net impact on aggregate savings. Since the argument for an historic aggregate reduction in savings invokes the life cycle theory and the wealth increment effect, the findings lend little support to the notion that Social Security has reduced the capital stock. The results cast doubt on the ability of people to accurately project their Social Security benefits and their age of retirement; large differences in lifetime wealth generated by the Social Security systems do not appear to influence savings. The expected age of retirement coefficient does, on the other hand, provide some support for the life cycle theory as opposed to a simple Keynesian consumption out of disposable income theory. The magnitude of the retirement coefficient is small; its size rules out the possibility that increased savings due to induced early retirement substantially offset the reduction in accumulation due to Social Security taxation (*ASST*). Indeed our parameterization of Social Security's implicit tax on post-62 work effort indicates that Social Security does not significantly influence the expected retirement decision.

In conclusion I feel that the final resolution of Social Security's effect on capital accumulation will require additional empirical work at both the micro and macro levels. Of particular interest at the micro level is determining the savings behavior of the post-65 age group. If every dollar of taxes reduces savings of the young by $.666, what does a dollar of benefits imply for the savings of the elderly? At the opposite end of the age distribution, the savings response of the very young to Social Security deserves attention, this age group is both more likely to face borrowing constraints and less likely to think about retirement when deciding how much to save. Additional empirical work as well as improved data detailing earnings histories and pension contributions will hopefully resolve the question of Social Security's effect on capital accumulation.

Appendix: Construction of Social Security Variables

Feldstein has presented concepts of gross and net Social Security wealth in a number of recent articles. The Social Security variables *ASST* and LWI_x presented here are closely related to his 1976 Social Security wealth variables, and the methodology used to construct these variables follows his. All computations to form *ASST* and LWI_x are based on separate information for the head and wife. For example, *ASST* is the sum of the head's accumulated taxes and that of the wife's where the head's accumulated taxes are based on his average (1965 and 1966) labor income and the wife's accumulated taxes are based on her average (1965 and 1966) labor income. Since earnings histories are not available, *ASST* was generated by applying historic growth rates of nominal wages to the two-year average of labor income (here gross of income taxes) as well as historic tax rates

and tax ceilings. The estimated paid in taxes were accumulated forward at a 6 percent nominal rate of interest.

The starting date for tax accumulation was 1937 or later. For respondents indicating they were in college beyond 1937, accumulation began the year they finished college. Other respondents indicated they had begun full-time work before age 20. In this case the year such work began or 1937 was used depending on which occurred latter. In all other cases the larger of 1937 and the year the respondent reached age 20 was the starting date. For wives, accumulation started the last year of college, the year they reached 20, or 1937 depending on which year was latest. The tax ceilings, starting age, marital status, employment of wife, income tax rate, and ages of the head and wife all contribute to variation in *ASST* independent of the household average labor income (*ALY*).

Turning to the LWI_x variable, the 1937 starting date for the Social Security system means that for the majority of heads and wives in my sample, a substantial number of their early working years were years of no Social Security tax payments. Since benefits are only loosely related to tax contributions LTW_x is a positive and large number for almost all households (see Table 1). The higher than interest yield on Social Security is due here not to a rate of labor force plus productivity growth in excess of the rate of interest, rather it is due to a truncation of the age-tax profile at the early end, that is, the intergenerational transfer.

In computing LWI_x, the present expected value of future taxes and future benefits are needed. To obtain the former of these magnitudes I project a real growth of current Social Security taxes at 2 percent up to age 65 and discount them back at a 3 percent real rate of interest applying survival probabilities at each age for the husband and wife separately.[13] The computation of present expected future benefits is much more involved. First basic benefits were assigned to the head and wife based on their individual current average labor income (again gross of income taxes). Next these basic benefits were assumed to grow at a 2 percent rate into the future and to become available for the head or wife the year the head or wife attains age 65.[14] Beyond the age 65 both the husband and the wife may be eligible for dependent or widow (widower) benefits. For each year the maximum possible benefit under each household survival contingency is used, the probability of each contingency occurring is applied to this benefit, and the expected benefit is discounted back. Thus, for example, if the wife can collect more as a dependent than she can collect based on her own tax contributions, this higher benefit is used for the contingencies that both spouses are alive. This assignment of highest possible benefits is the Social Security procedure.

Two methods of allocating basic benefits were tried. The first method treats individuals as if they could retire immediately and collect benefits based on their computed average monthly wage. The second method applies replacement ratios to current average labor income where the replacement ratios are based on the individual's position in the distribution of labor income. These replacement ratios are derived from looking at the distribution of benefits of new retirees. Since the results proved quite insensitive to which benefit allocation method was used, I report only results using the first method.

Finally, as mentioned above, the *NLS* data reports on Social Security eligibility. For heads reporting they would never be eligible for Social Security, all head components of household Social Security variables were set to zero and the wife was permitted no dependent or widow benefits.

[13]In this and other computations separate male and female survival probabilities were constructed from information provided in Social Security Administration (1966a).

[14]The early retirement provisions of Social Security permit the receipt of these benefits as of age 62 but on an actuarially reduced basis. This actuarial reduction makes our computation independent of whether age 62, 63, 64, or 65 is used as the starting age.

REFERENCES

J. D. Adams, "Estimates of Wealth Elasticity of Bequest," unpublished doctoral dissertation, Univ. Chicago 1975, ch. 3.

M. Boskin, "Social Security and Retirement Decisions," *Econ. Inquiry*, Jan. 1977, *15*,

1–25.

C. D. Campbell and R. G. Campbell, "Conflicting views on the Effect of Old-Age and Survivors Insurance on Retirement," *Econ. Inquiry*, Sept. 1976, *14*, 369–88.

M. Feldstein, "Social Security, Induced Retirement and Aggregate Capital Accumulation," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 905–26.

———, "Social Security and the Distribution of Wealth," *J. Amer. Statist. Assn.*, Dec. 1976, *71*, 800–07.

——— **and A. Pellechio,** "Social Security and Household Wealth Accumulation: New Microeconometric Evidence," disc. paper no. 530, Harvard Instit. Econ. Res., Jan. 1977.

J. Heckman, "Sample Selection Bias as a Specification Error," R-1984-HEW, Rand Corp., Feb. 1977.

L. J. Kotlikoff, "Social Security and Equilibrium Capital Intensity," *Quart. J. Econ.*, forthcoming.

P. Lauriat and W. Rabin, "Men Who Claim Benefits Before Age 65: Findings From the Survey of New Beneficiaries, 1968," *Soc. Sec. Bull.*, Nov. 1970, *33*, 3–28.

A. H. Munnell, "The Impact of Social Security on Personal Savings," *Nat. Tax J.*, Dec. 1974, *17*, 553–67.

———, Private Pensions and Saving: New Evidence," *J. Polit. Econ.*, Oct. 1976, *84*, 1013–32.

Manpower Report of the President, Washington 1975.

"National Longitudinal Survey 1966–69," Ohio State Univ.

Public Health Service, *Vital Statistics of the U.S., 1967*, Vol. 2, Part A, Mortality, Washington 1969.

Social Security Administration, (1966a) "U.S. Population Projections for OASDHI Cost Estimates," actuarial stud. no. 62, Office of the Actuary, Dec. 1966.

———, (1966b) *Social Security Handbook*, 3d ed., Washington 1966.

Job Search and Vacancy Contacts

By JOHN J. SEATER*

In this paper, I show the need for and then derive a model in which the number of vacancies contacted by a job searcher depends on the intensity of search, with diminishing returns to search intensity.

The usual assumption of the search literature (for example, see Dale Mortensen, 1970a) is that the number of vacancies contacted per period by a searcher is fixed. Such fixity is unsatisfactory for two obvious reasons. First, participants can vary the intensity of search and consequently can vary the number of vacancies contacted per period. Second, the number of existing vacancies changes over the business cycle so that, for a given search intensity, the number of vacancies contacted should change, too.

Besides these flaws, the assumption of a constant vacancy contact rate turns out to imply that a participant never chooses to work and search simultaneously. Simultaneous work and search requires that the function describing a searcher's vacancy contact rate exhibit diminishing returns to search intensity, a property I call diminishing physical returns to search. The requirement for diminishing physical returns holds even when the individual maximizes a concave utility function; diminishing utility returns to search are insufficient to allow simultaneous work and search. Thus, for example, Mortensen's (1977) discussion of simultaneous work and search is inappropriate in his utility-maximizing model in which the vacancy contact rate is a linear function of search intensity.

Even though simultaneous work and search requires diminishing physical returns to search, it is not immediately apparent what the source of such diminishing return would be.[1] At first glance, it may seem that if a searcher doubles his search intensity, he should double the number of vacancies he contacts. However, under the assumption that search requires travelling to firms to seek vacancies (rather than calling on the telephone, for example), it turns out that an increase in search intensity leads to a less than proportional increase in vacancy contacts. The reason is essentially that increasing the number of vacancies contacted requires an increase in the area over which search is conducted and the travel time required to cover this area increases faster than the area to be covered. Thus diminishing returns arise for spatial reasons.

The model developed in this paper proves the foregoing propositions and also provides some other interesting insights, the most important being a nullification of the criticism by James Tobin, Robert J. Gordon, and others that search theory produces incorrect predictions of the cyclical behavior of quits.

I. A General Search Model

In this section, I summarize a very general search model that contains most earlier search models as special cases; a more detailed discussion is in my 1977 paper. Using this model and a general vacancy contact function, I compare the results of assuming constant vs. diminishing physical returns to search intensity.

A. *Utility Function*

Assume the individual derives utility from consumption and leisure. Consumption at any

*Senior economist, department of research, Federal Reserve Bank of Philadelphia. I thank Robert A. Jones and George Borts for helpful comments on earlier drafts. The views expressed herein are not necessarily those of any branch of the Federal Reserve System.

[1]Steven Lippman and John McCall (1976b) and the author (1977) present analyses of search that depend crucially on diminishing physical returns to search intensity; however, in these contributions diminishing physical returns are neither derived nor motivated but merely assumed.

moment t, denoted $C(t)$, is nonnegative and is measured in goods per week. Leisure is all time not spent in labor and job search. The rate of leisure at any moment t, measured in hours per week, is $168 - L(t) - S(t)$, where 168 is the number of hours in a week and $L(t)$ and $S(t)$ are the rates of labor and search at moment t, measured in hours per week. Labor is time spent working on a job, and job search is time spent looking for a higher paying job than that currently held. The rates of labor and search, and also the sum of these rates, must be between 0 and 168.

The assumption that the individual derives utility from consumption and leisure implies that this utility depends only on C and the sum $L + S$ and not on the sizes of the individual terms L and S. Consequently, utility at time t can be expressed as $U[C(t), L(t) + S(t)]$. I assume that U is concave in consumption and leisure, is additively separable in C and $L + S$, and satisfies the usual Inada conditions; thus

$$U_1 > 0, U_{11} < 0$$
$$U_2 < 0, U_{22} < 0$$
$$U_{12} = U_{21} = 0$$
$$\lim_{C \to 0} U_1 = +\infty \quad \lim_{C \to \infty} U_1 = 0$$
$$\lim_{L+S \to 0} U_2 = 0 \quad \lim_{L+S \to 168} U_2 = -\infty$$

B. *The Budget Constraint*

The individual's lifetime budget constraint is

$$(1)\quad (d/dt)[A(t)/p] = r[A(t)/p] + [w(t)/p]L(t) - C(t)$$

$$(2)\quad A(T) = A_T$$

$$(3)\quad A(D) = 0$$

where

$A(t)$ = expected total assets at time t, measured in current dollars;

r = the expected interest rate at time t, assumed constant;

$w(t)$ = the expected nominal hourly wage rate at time t, measured in current dollars;

p = the expected price of consumption goods at time t, assumed constant;

T = the present moment;

D = the moment of death, assumed known by the individual;

A_T = the initial value of assets, known by the individual.

For simplicity, it is assumed that all assets have variable interest rates.

C. *Wage Improvement*

The individual has some control over the future value of his wage through job search. Assume that:

a) The individual has a perception of the distribution of nominal wages over vacancies, denoted F. Let f be the density function associated with F: $f = dF/dw$.

b) The individual's nominal wage will not change unless he searches.

c) The individual must accept or reject an offer as soon as he receives it.

d) Similarly, if the individual quits his job, he loses all rights to it.

e) There is no cost to changing jobs.

f) The only costs of search are foregone leisure and earnings.

At any moment t, the rate at which the individual contacts vacancies is a function N of the rate of search at moment t. Thus at any moment the rate of contact is $N[S(t)]$, which has the units vacancies per week. Other possible arguments of N such as the vacancy and unemployment rates or transportation technology are suppressed for now but will be discussed in Section II. The function N is a production function giving the physical returns to search, that is, vacancy contacts. Allow N to be either linear or concave in search: $N'(S) > 0$, $N''(S) = 0$. If N is concave in search, then assume that

$$\lim_{S \to 0} N'(S) = +\infty \quad \lim_{S \to +\infty} N'(S) = 0$$

These limit properties will be justified in Section II.

Suppose an individual earning wage $w(t)$ decides to look into a vacancy chosen at random. His estimate of the probability that he fails to improve his wage is $F[w(t)]$. Suppose the individual searches for a length

of time Δt, and suppose $N[S(t)]$ is constant over Δt. The total number of vacancies the individual can contact is $N[S(t)]\Delta t$. His estimate of the probability that he will fail to improve his wage during the entire period Δt is

$$(F[w(t)])^{N[S(t)]\Delta t}$$

The density function associated with $(F[w(t)])^{N[S(t)]\Delta t}$ is

$$(d/dw)F[w(t)]^{(N[S(t)])\Delta t}$$

which I denote by $f_{N,\Delta t}[w(t)]$. At any time t, the expected improvement in the wage over the period Δt, denoted $E[\Delta w(t)]$, is

$$E[\Delta w(t)] = \int_{w(t)}^{w^*}[\omega - w(t)]\, f_{N,\Delta t}(\omega)\, d\omega$$

where w^* is the maximum wage attainable. This function can be manipulated into the more manageable form

$$E[\Delta w(t)] = \int_{w(t)}^{w^*}(1 - [F(\omega)]^{N[S(t)]\Delta t})\, d\omega$$

The expected rate of change of $w(t)$ at time t, denoted $E[dw(t)/dt]$, is obtained by taking the appropriate limit as Δt goes to zero:

$$(4) \quad E[dw(t)/dt] = -N[S(t)] \int_{w(t)}^{w^*} ln\, F(\omega)\, d\omega \geq 0$$

I define

$$(5) \quad E[dw(t)/dt] \equiv g[w(t), S(t)]$$

The function g gives the revenue returns from search. It has the following derivatives:

$$g_w \leq 0,\ g_{ww} \geq 0,\ g_S \geq 0,\ g_{SS} \leq 0,\ g_{Sw} \leq 0,$$

These all are equalities if $w = w^*$; g_w and g_{ww} equal zero if $S = 0$; and g_{SS} equals zero if $N''(S) = 0$. Otherwise, the above derivatives all are nonzero.

Relation (5) is the individual's best estimate of the future rate of nominal wage improvement. I assume that the individual treats it as the rate that actually will occur, and write[2]

$$(d/dt)w(t) = g[w(t), S(t)]$$

The rate of change of real wages is simply

$$(6) \quad (d/dt)[w(t)/p] = g[w(t), S(t)]/p$$

which is the needed function relating wage improvement to job search.

D. *Utility Maximization*

The individual's utility over his remaining lifetime is[3]

$$(7) \quad \int_T^D U[C(t), L(t) + S(t)]dt$$

The individual's problem is to maximize (7) subject to his budget constraint given by (1), (2), and (3), and his real wage improvement constraint given by (6), together with the initial condition

$$(8) \quad w(T) = w_T$$

where w_T is known by the individual.[4]

As long as A is not too negative, as I will assume, a solution or optimal control exists for this problem. The optimal control can be characterized with Pontryagin's maximum principle. The Hamiltonian is

$$H = U[C(t), L(t) + S(t)] + \psi(t)\left[r\frac{A(t)}{p} + \frac{w(t)}{p}L(t) - C(t)\right] + \lambda(t)\frac{g[w(t), S(t)]}{p}$$

[2]This assumption implies certainty equivalence and ignores behavior toward risk. I discuss this simplification further in fn. 4.

[3]Insertion into (7) of the usual discount factor $e^{-\delta t}$ would not alter the results substantially. However, when the time horizon is known and finite, the discount rate neither is needed to guarantee existence of the integral nor has the natural interpretation that arises when the time horizon is infinite.

[4]This formulation, which maximizes utility as a function of expected wages, eliminates the stochastic elements of the problem and ignores risk. It would be better to treat this problem by maximizing expected utility; however, to do so would require solving a difficult stochastic control problem. I have simplified matters by solving a deterministic approximation instead. In my 1977 paper I discuss some of the stochastic aspects of the problem with an indirect method. For discussions of some of the risk considerations of the time allocation problem, see David Whipple, John Danforth, and David Levhari and Yoram Weiss.

If a control is optimal, then[5]

$$(9)\quad (d/dt)[A(t)/p] = \partial H/\partial\psi = r[A(t)/p] + [w(t)/p]L(t) - C(t)$$

$$(10)\quad (d/dt)[w(t)/p] = \partial H/\partial\lambda = g[w(t), S(t)]/p$$

$$(11)\quad (d/dt)\psi(t) = -\partial H/\partial(A/p) = -r\psi(t)$$

$$(12)\quad (d/dt)\lambda(t) = -\partial H/\partial(w/p) = -\psi(t)L(t) - \lambda(t)g_w[w(t), S(t)]$$

Equations (9) and (10) are simply restatements of (1) and (6). The adjoint variables ψ and λ are the marginal values of assets and wages, respectively, and equations (11) and (12) describe the motion of these variables over time. An optimal control also must satisfy initial conditions (2), (3), and (8); furthermore, by the transversality condition, it must satisfy

$$(13)\quad \lambda(D) = 0$$

Finally, an optimal control must satisfy the marginal conditions

$$(14)\quad \partial H/\partial C = U_C - \psi = 0$$

$$(15)\quad \partial H/\partial L = U_L + \psi(w/p) \leq 0$$

$$(16)\quad \partial H/\partial S = U_S + \lambda(g_S/p) \leq 0$$

When equality holds in (15) and (16), relations (14)–(16) simply express the usual equalities between marginal benefits and marginal costs. However, equality need not hold in (15) or (16), as I discuss momentarily.

[5]Conditions (9)–(16) are necessary for optimal control but may not be sufficient. A standard theorem states that if the Hamiltonian is concave in the state variables, then the necessary conditions are also sufficient. In the problem at hand, the Hamiltonian is convex in w, so the standard theorem does not apply. Fortunately, the concavity of the utility function U in consumption and leisure and the lack of dependence of the objective function (7) on the state variables A and w guarantee that a control for which (2), (3), and (8)–(15) hold is at least locally optimal. See E. Lee and L. Markus for details.

As explained in my 1977 paper, the foregoing model contains as special cases virtually all the earlier search models (and several human capital models, with appropriate changes) cited in the reference section.

E. *Alternative Marginal Returns to Search*

In finding the optimal control (i.e., the optimal paths for C, L, and S), expressions for C, L, and S as functions of ψ, λ, and w must be obtained. This is done by totally differentiating system (14)–(16) to obtain

$$(17)\quad \begin{bmatrix} dC \\ dL \\ dS \end{bmatrix} = \begin{bmatrix} U_{11} & 0 & 0 \\ 0 & U_{22} & U_{22} \\ 0 & U_{22} & U_{22} + \lambda g_{SS} \end{bmatrix}^{-1} \cdot \begin{bmatrix} -1 & 0 & 0 \\ -w/p & 0 & -w/p \\ 0 & -g_{S/p} & -g_{Sw} \end{bmatrix} \begin{bmatrix} d\psi \\ d\lambda \\ dw \end{bmatrix}$$

There are two cases to consider. First, suppose there are diminishing physical returns to search. Then $N''(S)$ is strictly negative so that, by (4) and (5), g_{SS} also is strictly negative. Consequently, the inverse matrix on the right side of (17) exists, and a solution to (17) is well-defined. The properties of this solution are discussed in detail in my 1977 paper, where it is shown that L and S behave as follows. Very early in the searcher's life, when w is near zero, (15) will be satisfied as an inequality and (16) as an equality. This is seen most clearly at the first moment the individual enters the labor force, when w equals zero. With w equal to zero, (15) could be satisfied by an equality only if L and S were both zero, which is ruled out by the limit conditions postulated at the end of Section IIA. Thus at least one of L or S must be positive, implying that U_2 is negative and therefore that (15) must be satisfied as an inequality. Because (15) is the marginal cost/benefit condition for determining labor supply, its being an inequality implies that L is set at zero. Thus S is positive, and its value is picked to satisfy (16) as an equality. The same sort of argument holds when w is positive but small. In the early part of his life, the individual searches but does not work. As time passes, w grows, and (15) tends toward

equality.[6] Once (15) becomes an equality, L becomes positive. Relation (16) remains an equality, and S continues to be positive throughout the individual's lifetime, falling to zero at exactly time D. Thus, except for an initial period when the individual searches but does not work, search and work occur simultaneously.[7]

Now consider the case of constant physical returns to search. Then $N''(S)$ equals zero, and so does g_{SS}. Consequently, the inverse matrix on the right side of (17) does not exist, and a solution to (17) is in general not defined. The reason for this lack of a solution can be seen from (15) and (16). Recall that $U_L = U_S = U_2$; and note that when $g_{SS} = 0$, the second terms of (15) and (16) are independent of C, L, and S. Thus a combination of C, L, and S must be found such that U_2 simultaneously equals two different constants: $-\psi w/p$ from (15) and $-\lambda g_S/p$ from (16). In general, such a simultaneous solution cannot be found, so that at all times one of (15) and (16) is satisfied as an inequality. The properties of the optimal control in this case are as follows. Initially, when w is small, L is set equal to zero and S is positive, just as in the case when $N''(S) = 0$. However, now when w rises to the point where (15) is satisfied as an equality and L becomes positive, (16) becomes an inequality and S is set equal to zero. Thus simultaneous work and search do not occur.

It seems, then, that if simultaneous work and search are to emerge, diminishing physical returns to search must be assumed.[8] Notice that this conclusion holds even though the individual maximizes a concave utility function and therefore experiences diminishing utility returns to search whether physical returns diminish or not.

II. The Vacancy Contact Function N

In the real world, we observe simultaneous work and search, so from the foregoing arguments it seems we should observe diminishing physical returns to search. At first glance, however, it is not obvious what the cause of diminishing returns would be. It might seem that an increase in the number of hours spent per week in search should increase equiproportionally the number of vacancies contacted per week. As shown in this section, spatial aspects of the search process can cause diminishing returns.

A. *A Spatial Model*

During any period, the rate at which a searching individual finds vacancies is a function of the rate of search during that period. Let X be the area of the region in which the individual conducts his search, τ the time required to travel one unit of distance, and V and B the total numbers of vacancies and firms (businesses). Suppose that the individual decides to visit in one period all firms within a circle of radius R around his house. Then, on average, the total number of firms visited in one period, N_B, will be

$$N_B = \pi R^2 \left(\frac{B}{X}\right) \tag{18}$$

which is the area of the circle multiplied by the average density of firms.

Suppose the individual searches by leaving his house, visiting one firm, and then returning to his house before visiting another firm. It takes 2τ hours to visit a firm one unit of distance away, so the total time spent searching all firms within the circle of radius R is

[6] The meaning of the wage growing when the individual is not working needs clarification. Consider an individual at initial time T formulating his optimal plan. Suppose he plans to commence work at time T* at an expected wage of $w(T^*)$. Consider a time $T' < T^*$. The wage $w(T')$ is the expected value of the wage the individual could earn if he followed the same pattern of search but commenced work at T′ instead of T*. Clearly, $w(T^*) > w(T')$. It is in this sense that the wage grows when the individual is searching but not working. See my 1977 paper for more discussion of this issue.

[7] My 1977 paper also discusses a modification of the model which allows both L and S to become zero later in the individual's life so that retirement is possible.

[8] It is possible that a complete analysis of risk also would allow simultaneous work and search. Risk-averse individuals might not want to put all their eggs in one basket. However, it also is possible that risk merely would change the duration and intensity of search. One still might obtain a bang-bang control unless diminishing physical returns to search were assumed. Intuition (at least mine) seems inadequate to settle this issue.

$$S = \sum_{i=1}^{N_B} 2\tau z_i$$

where z_i is the distance to the ith firm. To convert this expression to a more convenient form, first obtain a crude approximation by setting all the z_i equal to R, which gives

$$S \simeq \sum_{i=1}^{N_B} 2\tau R$$

Now refine this approximation by first partitioning the circle into m concentric circles of radii R_j, where $j = 1, 2, 3, \ldots, m$ and $R_m = R$, and then setting equal to R_j all the z_i for which $R_{j-1} < z_i \leq R_j$, where $R_0 = 0$. This gives

$$S \simeq \sum_{j=1}^{m} (2\tau R_j)\pi(R_j^2 - R_{j-1}^2)\frac{B}{X}$$
$$= \sum_{j=1}^{m} (2\tau R_j)\pi(R_j + R_{j-1})\frac{B}{X}(R_j - R_{j-1})$$

Taking the limit as m goes to infinity, this last expression equals S exactly:

$$S = \int_0^R (2\tau z)\pi 2z\frac{B}{X}dz = \frac{4\pi\tau BR^3}{3X}$$

Manipulation yields

$$R = \left(\frac{3XS}{4\pi\tau B}\right)^{1/3}$$

Therefore, from (18),

$$N_B = \pi\frac{B}{X}\left(\frac{3XS}{4\pi\tau B}\right)^{2/3}$$

which is the number of firms contacted per period. The number of vacancies N contacted per period is simply N_B multiplied by the average ratio of vacancies to firms:

$$\text{(19)} \qquad N(S) = N_B\left(\frac{V}{B}\right)$$
$$= \left(\frac{\pi}{X}\right)^{1/3} V\left(\frac{1}{B}\right)^{2/3}\left(\frac{3S}{4\tau}\right)^{2/3}$$

which is the desired function. The function is concave in search intensity:

$$N'(S) = \frac{1}{2\tau}\left(\frac{\pi}{X}\right)^{1/3} V\left(\frac{1}{B}\right)^{2/3}\left(\frac{3S}{4\tau}\right)^{-1/3} > 0$$

$$N''(S) = -\frac{1}{8\tau^2}\left(\frac{\pi}{X}\right)^{1/3} V\left(\frac{1}{B}\right)^{2/3}\left(\frac{3S}{4\tau}\right)^{-4/3} < 0$$

Note that $N''(S)$ has the limit properties assumed in Section Ic.

The cause of the diminishing physical returns to search intensity is spatial. Suppose an individual is searching a circle of radius R_1 and considers increasing this radius by exactly enough to double the number of firms contacted. Let this radius be R_2. The situation is depicted in Figure 1. The number of firms K contained in the inner circle equals the number contained in the annulus. Clearly, the total travel time required to search all firms in the inner circle is less than $2\tau R_1 K$, whereas the total travel time required to search all firms in the annulus is more than this. Thus doubling the number of firms searched more than doubles total search time; consequently, doubling search time would less than double vacancy contacts.

Diminishing returns to search intensity is a crucial property in the analyses of Lippman and McCall (1976b), and the author (1977). With it we obtain several interesting results on search behavior. However, these authors neither derive nor motivate diminishing returns but rather merely assume them. The model developed in this section provides a justification for this assumption.

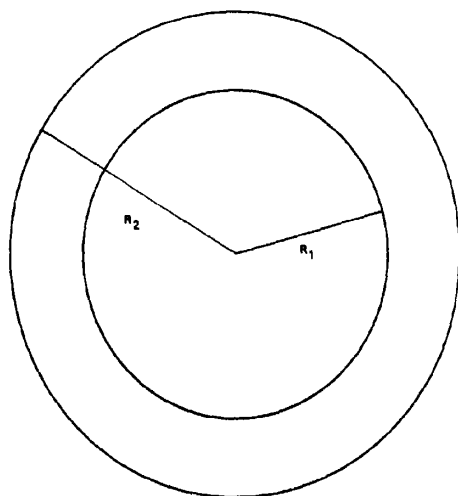



FIGURE 1

B. *Implications*

Besides justifying the assumption of diminishing physical returns to search and therefore providing an explanation for simultaneous work and search, the vacancy contact model just developed has several other interesting implications.

Perhaps the most important implication concerns a criticism of search-theoretic explanations of the Phillips curve. Such explanations argue that, when the economy slides into a recession and employers begin cutting wages, workers initially perceive these wage cuts as specific to themselves rather than as part of an economy-wide phenomenon. Workers thus believe that their relative wages have fallen, so they quit work to search for the better jobs they believe to exist. Their action raises unemployment and thus produces a Phillips curve. However, as Tobin, Gordon, and others have pointed out, quits do not rise in recessions but rather fall. Consequently, these critics argue, search-theoretic explanations of the Phillips curve must be viewed with a least some doubt.

The problem with search-theoretic quit predictions does not lie with search theory per se but rather with the way search theory usually is formulated. Virtually all search models assume the number of vacancies to be constant. In fact, of course, vacancies vary over the cycle, increasing in booms and decreasing in busts. As Tobin points out, quits depend not only on preceived wage differentials but also on the perceived probability of obtaining a job. As the theory developed above shows, this probability depends positively on the vacancy rate. Thus in a recession, when vacancy rates fall, the probability of landing a job decreases, tending to reduce quits and thus to offset the effect of increased perceived wage differentials. Which force dominates is an empirical question, and the evidence suggests that the reduced probability of finding a vacancy dominates the increased perceived wage differential.[9] In any case, when search theory is modified to account for the cyclical variability of vacancy rates, it no longer is invalidated by the empirical facts of quit behavior.

Of course, there remains the valid criticism of most search models that they look only at labor supply and not labor demand. Consequently, they cannot provide a full explanation of the equilibrium values of employment, quits, vacancies, etc. Mortensen (1970a) is the only author to attempt a full analysis of the labor market within a search framework. Even his demand side is flawed in several ways; for example, expectations are given, firms set wages but not prices, employment always is supply rather than demand determined. Moreover, the restrictions that are implied by Mortensen's demand side do not seem to alter the qualitative conclusions that emerge from the simpler one-sided models. Thus search theory still seems to be incomplete. However, the point of the foregoing argument is that the Tobin-Gordon criticism, which attempts to invalidate search theory *on its own ground*, is nullified by a more careful formulation of search theory.

Another implication of the above vacancy contact model is that equiproportional changes in the numbers of vacancies and firms do not leave N unchanged, but rather change N in the same direction as the changes in V and B. Again, the reason is spatial and is most easily seen from equation (19). An equiproportional increase, for example, in V and B leaves the ratio V/B unchanged. The ratio is the probability that a firm chosen at random has a vacancy. However, the increase in B makes it easier for the searcher to contact firms and therefore raises N_B. Thus N must rise. Empirically, this relation does not seem to be important for the *U.S.* economy as a whole because the variation in B is very small and, for the very few years when vacancy data

[9]Lippman and McCall (1976b) obtain the results that quits necessarily move procyclically. However, these results seems to be a consequence of their assumptions that participants always correctly perceive the nominal wage distribution, so that there are no misperceptions of relative wages, and that participants are income rather than utility maximizers and consequently ignore prices, so that there are no misperceptions of real wages. In the absence of these assumptions, quit behavior appears to be theoretically ambiguous. For an analysis of behavior by utility-maximizing participants who misperceive the nominal wage distribution, see my 1978 paper.

is available, is swamped by changes in V. Changes in B might be more significant relative to changes in V in some smaller areas such as cities or regions experiencing a flight of firms.

A third implication of the contact model is that the number of vacancy contacts is related inversely to travel time. Variations in travel time undoubtedly are significant over very long periods, such as the past seventy-five years (which includes years before and after the automobile). But changes can occur over much shorter spans of time, as with the introduction in urban areas of freeways and circumferential highways.

III. Summary

All job search literature recognizes the importance of the number of vacancies that a searching individual contacts per period; however, most search literature assumes this number to be constant. It seems more reasonable to assume that individuals have some control, through their search intensity, over the number of vacancies they contact per period. Using a fairly general model, I have shown how the optimal search intensity would be chosen; I also have shown that, unless the vacancy contact function exhibits diminishing returns to search intensity, simultaneous work and search does not occur. I then derived a simple vacancy contact function exhibiting diminishing returns to search intensity for spatial reasons. To increase the number of firms contacted, a searcher must increase the area over which he searches, which requires a more than proportional increase in travel time devoted to search. The vacancy contact function shows that contacts and hence quits are subject to procyclical forces through changes in the aggregate vacancy rate. These forces act in opposition to the usual wage misperception forces of search theory and render ambiguous the theoretical prediction of quit behavior's cyclical pattern. Thus, when modified by a variable vacancy contact rate, search theory no longer necessarily gives the wrong predictions for quit behavior that Tobin, Gordon, and others have criticized.

REFERENCES

R. D. Almonacid, "Nominal Income, Output, and Prices in the Short Run," unpublished doctoral dissertation, Univ. Chicago 1971.

Y. Ben-Porath, "The Production of Human Capital and the Life Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

A. S. Blinder and Y. Weiss, "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ.*, June 1976, *84*, 449–72.

J. P. Danforth, "Expected Utility, Mandatory Retirement and Job Search," disc. paper no. 74–141, Center Econ. Res., Univ Minnesota 1974.

M. Friedman, "The Role of Monetary Policy," *Amer. Econ. Rev.*, Mar. 1968, *58*, 1–17.

R. J. Gordon, "Recent Developments in the Theory of Inflation and Unemployment," *J. Monet. Econ.*, Apr. 1976, *2*, 185–220.

R. Gronau, "Information and Frictional Unemployment," *Amer. Econ. Rev.*, June 1971, *61*, 290–301.

J. Heckman, "Life Cycle Consumption and Labor Supply: An Explanation of the Relationship Between Income and Consumption Over the Life Cycle," *Amer. Econ. Rev*, Mar. 1974, *64*, 188–94.

E. B. Lee and L. Markus, *Foundations of Optimal Control Theory*, New York 1967.

D. Levhari and Y. Weiss, "The Effect of Risk on the Investment in Human Capital," *Amer. Econ. Rev.*, Dec. 1974, *64*, 950–63.

S. A. Lippman and J. S. McCall, (1976a) "The Economics of Job Search: A Survey: Part I," *Econ. Inquiry*, June 1976, *14*, 155–89; "Part II," Sept. 1976, *14*, 347–68.

______ and ______, (1976b) "Job Search in a Dynamic Economy," *J. Econ. Theory*, June 1976, *12*, 365–90.

R. E. Lucas, Jr. and L. A. Rapping, "Real Wages, Employment, and Inflation," in Edmund S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

J. J. McCall, "Economics of Information and Job Search," *Quart. J. Econ.*, Feb. 1970, *84*, 113–26.

D. T. Mortensen, (1970a) "A Theory of Wage

and Employment Dynamics," in Edmund S. Phelps, ed., *Microeconomic Foundations of Employment and Inflation Theory*, New York 1970.

———, (1970b) "Job Search, the Duration of Unemployment, and the Phillips Curve," *Amer. Econ. Rev.*, Dec. 1970, *60*, 847–62.

———, "Unemployment Insurance and Job Search Decisions," *Ind. Labor Relat. Rev.*, July 1977, *30*, 505–17.

S. **Rosen**, "Learning and Experience in the Labor Market," *J. Hum. Resources*, Summer 1972, *7*, 326–42.

J. J. **Seater**, "A Unified Model of Consumption, Labor Supply, and Job Search," *J. Econ. Theory*, Apr. 1977, *14*, 349–72.

———, "Utility Maximization, Aggregate Labor Force Behavior, and the Phillips Curve," *J. Monet. Econ.*, Nov. 1978, *4*, 687–713.

C. **Siven**, "Consumption, Supply of Labor and Search Activity in an Intertemporal Perspective," *Swedish J. Econ.*, Mar. 1974. *76*, 44–61.

G. **Stigler**, "Information in the Labor Market," *J. Polit. Econ.*, Oct. 1972, Part II, *72*, 94–105.

J. **Tobin**, "Inflation and Unemployment," *Amer. Econ. Rev.*, Mar. 1972, *62*, 1–18.

Y. **Weiss**, "On the Optimal Lifetime Pattern of Labor Supply," *Econ. J.*, Dec. 1972, *72*, 1293–315.

D. **Whipple**, "A Generalized Theory of Job Search," *J. Polit. Econ.*, Oct. 1973, *81*, 1170–188.

The Arrow-Lind Theorem in a Continuum Economy

By ROY GARDNER*

In their stimulating contribution to the theory of cost-benefit analysis, Kenneth Arrow and Robert Lind stated:

> ...when the risks associated with a public investment are publicly borne, the total cost of risk-bearing is insignificant and, therefore, the government should ignore uncertainty in evaluating public investments.... This result is obtained not because the government is able to pool investments but because the government distributes the risk associated with any investment among a large number of people. It is the risk-spreading aspect of government investment that is essential to this result. [p. 366]

Arrow and Lind proved their result from an asymptotic point of view, letting the number of taxpayers grow infinitely large. Private national income grew large with the number of taxpayers, while the risk from public investment facing an individual taxpayer grew very small. Indeed, in the limit, the impact of the government investment relative to the private economy was negligible. With so many things happening at once, it was not altogether clear that the number of taxpayers alone was responsible for their striking result, although Arrow and Lind so argued (p. 373). In an attempt to clarify the situation, an alternative approach, that of an economy with a continuum of agents, is adopted here.[1] Such a model enables one to consider an economy with a fixed, very large number of agents, at the same time allowing the impact of risky government investment to vary. The major result for such a model is that the Arrow-Lind theorem holds if the government risk is *small* relative to the economy, a small risk being defined as one whose variance is arbitrarily small. A counterexample shows that if the government risk is not small, the Arrow-Lind theorem no longer holds, even though there are a great many agents over whom to spread the risk.[2] These results decide the issue: it is only for small government risks that the total cost of risk bearing is necessarily insignificant.

I. The Model

What follows adheres as closely as possible to the definitions and notation of Arrow and Lind. Consider then an economy whose agents, called taxpayers, belong to the set $T = [0, 1]$. The assumption that the economy consists of many small agents is expressed by taking an individual agent to be an infinitesimal subset dt of T. Since it is convenient to label agents by points t in T, one may also consider t as a typical point in the small set dt.

Each taxpayer t, $0 \leq t \leq 1$, has a twice-differentiable and bounded utility function U_t of income $Y(t)$. Its derivatives are denoted U'_t, U''_t. Since income Y can be random, each taxpayer is assumed to maximize expected utility, $EU_t(Y(t))$. Assuming boundedness avoids paradoxes of the St. Petersburg variety. It is further assumed that the marginal utility of income is positive for all taxpayers and all values of income.

Taxpayer income $Y(t)$ for any taxpayer t is given by

$$(1) \quad Y(t) = A(t) + s(t)\,[B^* + X]$$

where $A(t)$ is private income, a random variable, $s(t)$ is the distribution of government

*Assistant professor, Iowa State University. I wish to thank H. Lapan and an anonymous referee for their helpful comments.

[1] Robert Aumann and Lloyd Shapley (ch. 6) discuss the relationship between the two approaches and various interpretations of the latter.

[2] Results stressing other stringent conditions under which the Arrow-Lind theorem holds have recently been achieved by L. P. Foldes and R. Rees.

return, and B^* is the mean government return on investment. The random component of government return, X, is assumed to have mean zero and variance σ_X^2, and to be distributed independently of $A(t)$. Under these assumptions, expected taxpayer income is given by

$$(2)\quad EY(t) = \int_{A_0(t)}^{A_1(t)} \int_{X_0}^{X_1} A(t) + s(t) \cdot [B^* + X]\, dG(X)\, dF_t(A(t)) = \int_{A_0(t)}^{A_1(t)} A(t)\, dF_t(A(t)) + s(t)B^*$$

where dF_t and dG are the probability density functions of $A(t)$ and X, respectively. The distribution of government return satisfies

$$(3)\qquad 1 = \int_0^1 s(t)\, dt,\ s(t) > 0$$

where $s(t)$ is continuous. The government budget equation is

$$(4)\qquad \int_0^1 \int_{X_0}^{X_1} s(t)\, [B + X]\, dG(X)\, dt = B^*$$

Finally denoting by $EA(t)$ mean private income for taxpayer t, the mean national income is

$$(5)\qquad \int_0^1 EY(t)\, dt = \int_0^1 EA(t)\, dt + B^*$$

The risk premium $\pi(t)$ for taxpayer t satisfies by definition[3]

$$(6)\quad EU_t(Y(t)) = EU_t(A(t) + s(t)B^* - \pi(t))$$

Equation (6) expresses indifference on the part of taxpayer t to bearing his share $s(t)X\,dt$ of the public risk or paying $\pi(t)\,dt$ out of his income and bearing only private risk. For a risk averter, risk aversion and the risk premium are both positive. In this case, the risk premium represents a cost, namely the cost to taxpayer t of bearing the risk of the government investment. The total cost of government risk bearing is given by the integral

$$(7)\qquad \int_0^1 \pi(t)\, dt$$

[3]See John Pratt, p. 124. My $\pi(t)$ corresponds to Arrow and Lind's $k(n)$.

II. Results

One can now state the following analogue of the Arrow-Lind theorem for a continuum economy.

THEOREM: *Let the economy have a continuum of agents. Then the total cost of bearing government risk tends to zero, if the variance of the government risk tends to zero.*

PROOF:

One begins by estimating the risk premium for an individual agent.

From (1) and (6), it follows that

$$(8)\quad \int_{A_0(t)}^{A_1(t)} \int_{X_0}^{X_1} U_t(A(t) + s(t)[B^* + X]\, dG(X)\, dF_t(A(t)) = \int_{A_0(t)}^{A_1(t)} U_t(A(t) + s(t)B^* - \pi(t))\, dF_t(A(t))$$

Performing a Taylor-series expansion on the left-hand side yields[4]

$$(9)\quad \int_{A_0(t)}^{A_1(t)} \int_{X_0}^{X_1} U_t(A(t) + s(t)B^* + s(t)X\, U_t'(A + s(t)B^*) + \frac{s^2(t)\,X^2}{2} U_t''(A(t) + s(t)B^*) + o(\sigma_X^2)\, dG(X)\, dF_t(A(t))$$

while expansion of the right-hand side yields

$$(10)\quad \int_{A_0(t)}^{A_1(t)} \int_{X_0}^{X_1} U_t(A(t) + s(t)B^*) - \pi(t)U_t'(A(t) + s(t)B^*) + O(\pi^2(t))\, dG(X)\, dF_t(A(t))$$

Equating these expressions as (6) requires, one obtains after some manipulation

$$(11)\quad \pi(t) = \left[\frac{s^2(t)\sigma_X^2}{2} \int_A^{A_1(t)} U_t''(A(t) + s(t)B^*)\, dF_t(A(t))\right] \div \left[\int_{A_0(t)}^{A_1(t)} U_t'(A(t) + s(t)B^*)\, dF_t(A(t))\right] + o(\sigma_X^2)$$

[4]Pratt, p. 125, discusses the regularity conditions for the Taylor-series expansion. In the expansions, O ( ) means "terms of + order at most" and o ( ) means "terms of smaller order than."

Integrating (11) with respect to t as (9) requires, $\int_0^1 \pi(t)dt$ approaches zero as σ_X^2 approaches zero.

One may also observe that the total cost of government risk bearing is again small if the integral

$$(12)\quad \int_0^1 \int_{A_0(t)}^{A_1(t)} U_t''(A(t) + s(t)B^*)\, dF(A(t))\, dt$$

approaches zero. One may describe such a situation as one where the taxpayers are on average risk neutral.

When government risk is not small, (11) no longer applies. In this case one can show that even with a continuum of agents, each bearing an infinitesimal share of the risk, the total cost of risk bearing is not insignificant.

To this end, consider an economy of identical taxpayers:

$$(13)\quad U_t(Y(t)) = -Y^{-1}$$
$$s(t) = 1 \text{ for all } t$$
$$A(t) = A, \text{ a constant}$$

The public investment has positive value B^*, with risky component $X = 1$ with probability 1/2 and $X = -1$ with probability 1/2. Equation (8) in this case becomes

$$(14)\quad \frac{1}{2}\left(\frac{-1}{A + B^* + 1}\right) + \frac{1}{2}\left(\frac{-1}{A + B^* - 1}\right) = \frac{-1}{A + B^* - \pi(t)}$$

Solving for the risk premium,

$$(15)\quad \pi(t) = \frac{1}{A + B^*}$$

The total cost of government risk-bearing therefore is

$$(16)\quad \int_0^1 \pi(t)dt = 1/(A + B^*) > 0$$

The cost will approach zero if A grows large, which is the analogue of the Arrow-Lind argument in the continuum model.

Now suppose the government adopts this project. One can now reinterpret the situation as one of random private income.[5] The private income now is $A(t) = A + 1$ with probability 1/2 and $A - 1$ with probability 1/2. Let the government consider a second project, identical to the first, but distributed independently of it. One can show that the total cost of bearing the risk of this second project is again positive. For example, when $A = 1$ and $B^* = 2$, the cost of government risk bearing for the first project is .333, while the cost of government risk bearing for the second project is .463, not only positive but even larger.

III. Conclusion

Arrow and Lind's conclusion to ignore the uncertainty associated with public investment was the consequence of a limit argument, in which the number of economic agents became larger. At the same time, since each agent represented an addition to national income, the size of the government risk became small relative to the economy. It was a priori possible that one or both of these factors was responsible for the Arrow-Lind result, although Arrow and Lind argued exclusively in terms of the former. What has been shown is that for a continuum economy the decisive role is played by the latter factor. Benefit-cost analysis safely ignores the uncertainty associated with public investment when that investment's risk is inherently small.

[5] I am indebted to an anonymous referee for this observation.

REFERENCES

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

Robert J. Aumann and Lloyd S. Shapley, *Values of Non-Atomic Games*, Princeton 1974.

L. P. Foldes and R. Rees, "A Note on the Arrow-Lind Theorem," *Amer. Econ. Rev.*, Mar. 1977, *67*, 188–93.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

On the Equivalence of Input and Output Market Marshallian Surplus Measures

By STEPHEN E. JACOBSEN*

In a series of papers which has appeared in this *Review* the following question is raised: Given a change in the vector of unit factor prices, to what is the resulting change in Marshallian surplus (in the factor markets) equal? As is probably now well-known, the first two papers in this series, written by Richard Schmalensee (1971) and Daniel Wisecarver, answered the above question incorrectly. Indeed, James Anderson and Schmalensee (1976), in two separate papers, discussed the error of the first two papers. Anderson answers the question correctly under the assumptions of linear homogeneous and single output production, competition, and a change in only one factor price. Schmalensee corrects his previous paper by correctly answering the question under the assumptions of general single output production, competition, and a change in only one factor price. However, Schmalensee does not completely answer the question under monopolistic conditions and, as a result, appears to believe that it may be possible to infer Marshallian surplus change in the final output market from Marshallian surplus change in the factor markets.

The basic purpose of this article is to demonstrate the answers for both the monopoly and competitive cases under conditions more general than those of the above authors. In particular, the answers are derivable regardless of how many factor prices are changed and regardless of the number of final outputs which these factors produce. The approach employed in the last two papers cannot provide this generalization.

I will show that, under monopoly conditions, the change in Marshallian surplus in the factor markets (due to a factor-price vector change) is precisely equal to the change in monopoly profits that occurs because of the factor-price vector change. Also, it will be shown that under competitive conditions the Marshallian surplus change in the factor markets is precisely equal to the change in Marshallian surplus (measured in the final output markets) plus the change in producer's surplus (i.e., industry profits) which is the correct answer arrived at by both Anderson and Schmalensee in his second paper under the above mentioned restrictive conditions.

The generality of the answers is demonstrated by utilizing (i) what W. Erwin Diewert (p. 112) has called "Shephard's Lemma" (for example, see Paul A. Samuelson, p. 68; Ronald W. Shephard, p. 11; or Daniel McFadden p. 5), and (ii) the profit function from which the factor demands are derived.

I. The Approach

Let $u = (u_1, \ldots, u_m)$ be the levels of m outputs which are produced by the usage of n factors of production whose levels are denoted by the vector $x = (x_1, \ldots, x_n)$. Let $p = (p_1, \ldots, p_n)$ denote the vector of unit factor prices and assume that $p \geq 0$. Let $L(u)$ denote the level set for producing u. That is, $L(u)$ denotes the set of factor vectors x that are capable of producing the output vector u. The cost function is then defined as

$$(1) \qquad C(u, p) = \min\{px \mid x \in L(u)\}$$

and it is assumed that the minimum is attained by a vector in $L(u)$, denoted by $x(u, p)$; i.e.,

$$(2) \qquad x(u, p) = (x_1(u, p), \ldots, x_n(u, p))$$

is the optimal factor mix for producing the output vector u when faced with the factor-

*Associate professor, department of system science, School of Engineering and Applied Science, University of California-Los Angeles. This research was supported by the National Science Foundation, Grant No. ENG 76-12250

price vector p. Shephard's Lemma is

$$(3) \qquad \frac{\partial C(u,p)}{\partial p_i} = x_i(u,p) \qquad i = 1, \ldots, n$$

Now, let $r = (r_1, \ldots, r_m)$ denote the vector of unit output prices and let $D_i(r_1, \ldots, r_m)$, $i = 1, \ldots, m$, be the quantity demanded of the *ith* output when the output prices are given by $r_1, r_2, \ldots, r_m$; i.e., $u_i = D_i(r_1, \ldots, r_m)$, $i = 1, \ldots, m$. For simplicity, assume that these demand functions are invertible so that we may write $r_i = D_i^{-1}(u_1, \ldots, u_m)$, $i = 1, \ldots, m$. Let $R(u,p)$ denote the net revenue function; i.e.,

$$(4) \qquad R(u,p) = \sum_{i=1}^{m} D_i^{-1}(u)u_i - C(u,p)$$

Define the profit function (as a function of the factor-price vector) to be

$$(5) \qquad \pi(p) = \max_{u \geq 0} R(u,p)$$

That is, when faced with a factor-price vector p, the firm computes its optimal net revenue (i.e., its profits) by solving (5) and producing optimal output levels $u(p) = (u_1(p), u_2(p), \ldots, u_m(p))$, and therefore

$$(6) \qquad \pi(p) = R(u(p), p)$$

Now, the output levels $u_1(p), \ldots, u_m(p)$ are produced by an optimal factor mix

$$(7) \quad x(u(p),p) = (x_1(u(p),p), \ldots, x_n(u(p),p))$$

which I abbreviate as $x(p) = x(u(p),p)$. That is, $x(p) = (x_1(p), \ldots, x_n(p))$ is the optimal factor mix used to produce the optimal output levels $u_1(p), \ldots, u_m(p)$ when the factor-price vector is p.

The basic question, discussed in the introduction, can now be restated as follows. Let $p^0 = (p_1^0, \ldots, p_n^0)$ and $p^1 = (p_1^1, \ldots, p_n^1)$ be two factor-price vectors. We are interested in the evaluation of the line integral

$$(8) \qquad -\int_{p^0}^{p^1} x(p)\, dp$$

If this line integral is independent of the path of integration chosen when going from factor-price vector p^0 to factor-price vector p^1 (i.e., the value of the line integral is the same regardless of the path of integration), then (8) is unambiguously the change in Marshallian surplus in the factor markets, due to the change in the vector of factor prices. The reasons for the use of a line integral when generalizing surplus to several dimensions are (i) each of the factor demand functions $x_i(p)$, $i = 1, \ldots, n$, depends upon all factor prices, and (ii) we generally want, as shall be seen, that the integrand be the vector of partial derivatives of the integral (i.e., a generalization of the fundamental theorem of calculus) and this holds if, and only if, the value of the integral is independent of the path of integration. The use of line integrals for multidimensional surplus dates back to Harold Hotelling. When dealing with consumer derived demand functions, Eugene Silberberg has shown that path independence occurs if, and only if, the consumer's utility function is homothetic. I will show below that the corresponding situation for derived factor-demand functions is somewhat more straightforward.

In particular, if it can be shown that the vector $x(p) = (x_1(p), \ldots, x_n(p))$ is the vector of partial derivatives of some function of the vector p, say $f(p)$, then the independence of path property holds and so does the fundamental theorem of calculus. That is, we would then have, along any path of integration, that (for example, see Tom Apostol, p. 280)

$$(9) \quad \int_{p^0}^{p^1} x(p)\, dp = f(p^1) - f(p^0)$$

and

$$(10) \qquad \frac{\partial f(p)}{\partial p_k} = x_k(p), \;\; k = 1, \ldots, n$$

It turns out that the profit function $\pi(p)$ is extremely useful for this purpose. Since (4) and (6) imply

$$(11) \quad \pi(p) = \sum_{i=1}^{m} D_i^{-1}(u(p))u_i(p) - C(u(p),p)$$

we differentiate $\pi(p)$ with respect to p_k to obtain

$$(12)\quad \frac{\partial \pi(p)}{\partial p_k} = \sum_{i=1}^{m}\Big[\sum_{j=1}^{m} u_i(p) \frac{\partial D_i^{-1}(u(p))}{\partial u_j}\frac{\partial u_j(p)}{\partial p_k} + D_i^{-1}(u(p)) \frac{\partial u_i(p)}{\partial p_k}\Big] - \sum_{j=1}^{m}\frac{\partial C(u(p),p)}{\partial u_j}\frac{\partial u_j(p)}{\partial p_k} - \frac{\partial C(u(p),p)}{\partial p_k}$$

Rearranging, we obtain

$$(13)\quad \frac{\partial \pi(p)}{\partial p_k} = \sum_{j=1}^{m}\Big[\sum_{i=1}^{m} u_i(p) \frac{\partial D_i^{-1}(u(p))}{\partial u_j} + D_j^{-1}(u(p)) - \frac{\partial C(u(p),p)}{\partial u_j}\Big] \frac{\partial u_j(p)}{\partial p_k} - \frac{\partial C(u(p),p)}{\partial p_k},\quad k=1,\ldots,n$$

Case 1: *Monopoly*. Under monopoly conditions ($MC = MR$) we have that

$$(14)\quad \sum_{i=1}^{m} u_i(p)\frac{\partial D_i^{-1}(u(p))}{\partial u_j} + D_j^{-1}(u(p)) = \frac{\partial C(u(p),p)}{\partial u_j}$$

since the left-hand side is precisely marginal revenue (associated with the jth output) and the right-hand side is the marginal cost of the jth output. Both sides are evaluated at the optimal output levels $u(p) = (u_1(p),\ldots,u_m(p))$ and therefore the equality in (14) must hold. Therefore, (13) becomes

$$(15)\quad \frac{\partial \pi(p)}{\partial p_k} = -\frac{\partial C(u(p),p)}{\partial p_k} = -x_k(p)\qquad k=1,\ldots,n$$

where the second equality holds because of (3), Shephard's Lemma. It can be seen that the vector of derived factor demands $x(p) = (x_1(p),\ldots,x_n(p))$ is the vector of partial derivatives of the negative of the profit function. We can therefore immediately write (because of (9) and (10))

$$(16)\quad -\int_{p^0}^{p^1} x(p)\,dp = \pi(p^1) - \pi(p^0)$$

along any path of integration.

Equation (15) and hence (16) establish the fact that the factor-market change in Marshallian surplus under monopoly conditions is precisely equal to the change in monopoly profits that occurs because of a factor-price vector change from p^0 to p^1. Also, as we have shown, this result holds for changes in several factor prices and for multiple output production. Moreover, (16) shows that one can infer little if anything about the change in output-market surplus from the change in factor-market surplus under monopoly conditions.

Case 2: *Competition*. Let us idealize the competitive case by treating the industry as a monopoly which must choose output levels according to $MC = AR$. In particular, the optimization in (5) is carried out subject to this additional requirement, and therefore the values of $\pi(p)$, $u(p)$, and $x(p)$ are not the same as those of the monopoly case. Note that

$$(17)\quad D_j^{-1}(u(p)) = \frac{\partial C(u(p),p)}{\partial u_j}\qquad j=1,\ldots,m$$

(i.e., price equals marginal cost) and therefore (13) becomes, for $k = 1,\ldots,n$,

$$(18)\quad \frac{\partial \pi(p)}{\partial p_k} = \sum_{j=1}^{m}\sum_{i=1}^{m} u_i(p)\frac{\partial D_i^{-1}(u(p))}{\partial u_j}\frac{\partial u_j(p)}{\partial p_k} - x_k(p)$$

because of Shephard's Lemma and (17). Therefore, for $k = 1,\ldots,n$,

$$(19)\quad -x_k(p) = -\sum_{j=1}^{m}\sum_{i=1}^{m} u_i(p)\frac{\partial D_i^{-1}(u(p))}{\partial u_j}\frac{\partial u_j(p)}{\partial p_k} + \frac{\partial \pi(p)}{\partial p_k}$$

We can answer the basic question, as in the monopoly case, if we can determine whether the right-hand side of (19) is, for $k = 1,\ldots,n$, the kth partial derivative of some function. Toward this goal, first assume that there is only one output u_1 and its inverse

demand function is D_1^{-1}. Then, (19) becomes

$$(20)\quad -x_k(p) = -u_1(p)\frac{dD_1^{-1}(u_1(p))}{du_1}\frac{\partial u_1(p)}{\partial p_k} + \frac{\partial \pi(p)}{\partial p_k},\ k = 1,\ldots,n$$

The change in consumer's surplus associated with output levels $u_1(p^0)$ and $u_1(p)$ is

$$(21)\quad S_1(p;p^0) = -\int_{r_1(p^0)}^{r_1(p)} D_1(r_1)\,dr_1$$

where $r_1(p^0) = D_1^{-1}(u_1(p^0))$ and $r_1(p) = D_1^{-1}(u(p))$. We see that

$$(22)\quad \frac{\partial S_1(p;p^0)}{\partial p_k} = -u_1(p)\frac{dD_1^{-1}(u_1(p))}{du_1}\frac{\partial u_1(p)}{\partial p_k},\ k = 1,\ldots,n$$

Equations (20) and (22) therefore show that $-x_k(p)$, the negative of the kth factor derived demand function, is the kth partial derivative (with respect to p_k) of the function $S_1(p;p^0) + \pi(p)$. Therefore, as in the monopoly case, we can write, by using (9) and (10),

$$(23)\quad -\int_{p^0}^{p^1} x(p)\,dp = S_1(p^1;p^0) + \pi(p^1) - \pi(p^0)$$

along any path of integration (for the line integral on the left-hand side). The right-hand side of (23) is precisely the change in total surplus that occurs due to the price vector change from p^0 to p^1. Equation (23) generalizes Schmalensee's (1976) result to the case of several factor-price changes and one final output.

Now assume there are m outputs, $u_1,\ldots,u_m$, but their demand functions are independent. In particular, we assume $r_i = D_i^{-1}(u_i)$, $i = 1,\ldots,m$; that is, each output's quantity demanded depends only upon its own price. In this case, the change in consumer's surplus, associated with output levels $u(p^0) = (u_1(p^0),\ldots,u_m(p^0))$ and $u(p) = (u_1(p),\ldots,u_m(p))$ is

$$(24)\quad S(p;p^0) = -\sum_{i=1}^{m}\int_{r_i(p^0)}^{r_i(p)} D_i(\bar{r}_i)\,dr_i$$

where $r_i(p^0) = D_i^{-1}(u_i(p^0))$ and $r_i(p) = D_i^{-1}(u_i(p))$. We see that

$$(25)\quad \frac{\partial S(p;p^0)}{\partial p_k} = -\sum_{i=1}^{m} u_i(p)\frac{dD_i^{-1}(u_i(p))}{du_i}\frac{\partial u_i(p)}{\partial p_k}$$

The right-hand side of (25) is precisely the same as the first term on the right-hand side of (19) since, in this case, $\partial D_i^{-1}(u_i)/\partial u_j = 0$ for $i \neq j$. So, again, we have that $-x_k(p)$ for $k = 1,\ldots,n$ is the kth partial derivative with respect to p_k of the function $S(p;p^0) + \pi(p)$. Therefore, we have (because of (9) and (10))

$$(26)\quad -\int_{p^0}^{p^1} x(p)\,dp = S(p^1;p^0) + \pi(p^1) - \pi(p^0)$$

along any path of integration (for the line integral on the left-hand side). And, as before, the right-hand side is the change in total surplus that occurs due to the price-vector change from p^0 to p^1.

Under an additional assumption, it can be shown that (26) holds when the demand functions are jointly dependent. As before, let $r = (r_1,\ldots,r_m)$ and let $D_i(r)$, $i = 1,\ldots,m$, be the quantity demanded for the ith output when the output price vector is r. Let $u = (u_1,\ldots,u_m)$ and let $r_i = D_i^{-1}(u)$, $i = 1,\ldots,m$, be the inverse demand function for the ith output. Let $D(r) = (D_1(r),\ldots,D_m(r))$ be the vector of output demand functions and let $D^{-1}(u) = (D_1^{-1}(u),\ldots,D_m^{-1}(u))$ be the associated vector of inverse demand functions. In this notation we have $r(p) = D^{-1}(u(p))$.

If the line integral

$$(27)\quad -\int_{r^0}^{r} D(v)\,dv$$

has the independence of path property then this integral is unambiguously the change in consumer's surplus, in the output markets, due to an output-price vector change from r^0 to r. The importance of the independence of path property derives from the fact that if it did not hold, the value of the integral in (27) would vary depending upon the path of integration chosen. In particular, if the indepen-

dence of path property (which is equivalent to the "integrability conditions" $\partial D_i(r)/\partial r_j = \partial D_j(r)/\partial r_i$) does not hold, then consumer's surplus defined by (27) is ambiguous (for example, see Silberberg). However, Robert Willig has shown that this ambiguity is often insignificant since he has derived conditions under which (27), along a particular path, approximates the equivalent variation which is an important welfare measure for a consumer.

We make the additional assumption that the line integral in (27) is independent of the path of integration chosen. In this case, the change in consumer's surplus, associated with output levels $u(p^0) = (u_1(p^0), \ldots, u_m(p^0))$ and $u(p) = (u_1(p), \ldots, u_m(p))$ is

$$(28) \qquad S(p; p^0) = -\int_{r(p^0)}^{r(p)} D(r)\, dr$$

where $r(p^0) = D^{-1}(u(p^0))$ and $r(p) = D^{-1}(u(p))$. Under the path independence assumption, it can be shown[1] that

$$(29) \quad \frac{\partial S(p; p^0)}{\partial p_k} = -\sum_{j=1}^{m}\sum_{i=1}^{m} u_i(p)\frac{\partial D_i^{-1}(u(p))}{\partial u_j}\frac{\partial u_j(p)}{\partial p_k}$$

The right-hand side of (29) is precisely the first term on the right-hand side of (19). Therefore, again, we have that $-x_k(p)$ is the kth partial derivative of the function $S(p; p^0) + \pi(p)$. Again, because of (9) and (10),

$$(30) \quad -\int_{p^0}^{p^1} x(p)\, dp = S(p^1; p^0) + \pi(p^1) - \pi(p^0)$$

along any path of integration. As before, the right-hand side of (30) is the change in total surplus that occurs due to the factor-price vector change from p^0 to p^1.

[1]In fact, let $\nabla_p S(p; p^0)$ be the vector of partial derivatives, with respect to p, for the function $S(p; p^0)$. Then, since path independence holds by assumption, we have $\nabla_p S(p; p^0) = -u(p)^T \nabla_u D^{-1}(u(p)) \nabla u(p)$ where $u(p)^T$ is the transpose of the vector $u(p)$, $\nabla_u D^{-1}(u(p))$ is the Jacobian of the inverse demand vector, evaluated at $u(p)$, and $\nabla u(p)$ is the Jacobian of the output vector evaluated at p (for example, see Apostol). The kth component of $\nabla_p S(p; p^0)$ is the first term on the right-hand side of (19).

II. Conclusion

By explicitly considering the profit function, as a function of the factor-price vector, we arrive at equation (13). By then considering the two pricing rules (i.e., $MC = MR$ and $MC = AR$) for monopolistic and competitive production and then taking full advantage of Shephard's Lemma we arrive at the basic equations (15) and (19).

Equation (15) states that $-x(p)$, the negative of the factor derived demand vector, is the vector of partial derivatives of the profit function $\pi(p)$ when output is monopolistically produced. By the basic line integral theorem we can then write equation (16) which states that the change in Marshallian surplus in the factor markets is precisely equal to the change in monopoly profits that occurs because of the factor-price vector change. This result holds regardless of the number of changed factor prices and regardless of the number of final outputs.

When output demand functions are independent, or when dependent and the path independence property holds in the output markets, equation (19) states that $-x(p)$ is the vector of partial derivatives of the function $S(p; p^0) + \pi(p)$. By use of the basic line integral theorem we can then write equation (26) which states that under competitive output conditions, the change in Marshallian surplus in the factor markets is precisely equal to the change in total surplus that occurs due to the factor-price vector change.

An interesting by-product is that the joint derived factor-demand functions, $x_1(p), \ldots, x_n(p)$, usually satisfy the integrability conditions (i.e., path independence), a result which generally does not hold for a consumer's demand functions derived from utility maximization. In particular, the derived factor-demand functions are integrable under monopolistic output production and under competitive output production when there is path independence in the output markets.

REFERENCES

J. Anderson, "The Social Cost of Input Distortions: A Comment and a Generalization,"

Amer. Econ. Rev., Mar. 1976, *66*, 235–38.

Tom M. Apostol, *Mathematical Analysis*, 3d ed., Reading, Dec. 1960.

W. Diewert, "Applications of Duality Theory," in Michael D. Intriligator and David A. Kendrick, eds., *Frontiers of Quantitative Economics*, Vol. 2, Amsterdam 1974, 106–71.

H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica*, July 1938, *6*, 242-69.

D. McFadden, "Cost, Revenue, and Profit Functions: A Cursory Review," work. paper no. 86, Instit. Bus. Econ. Res., Univ. California-Berkeley 1966.

Paul A. Samuelson, *Foundations of Economic Analysis*, 6th ed., Cambridge, Mass. 1961.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.*, Sept. 1971, *61*, 682–87.

———, "Another Look at the Social Valuation of Input Price Changes," *Amer. Econ. Rev.*, Mar. 1976, *66*, 239–43.

Ronald W. Shephard, *Cost and Production Functions*, Princeton 1953.

E. Silberberg, "Duality and the Many Consumer's Surpluses," *Amer. Econ. Rev.*, Dec. 1972, *62*, 942–52.

R. Willig, "Consumer's Surplus: A Rigorous Cookbook," tech. rept. no. 98, IMSSS, Stanford Univ. May 1973.

D. Wisecarver, The Social Costs of Input-Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–72.

Inflation Expectations in the Monetarist Black Box

By SHLOMO MAITAL*

Jerome Stein (1974, 1976) boils the monetarist-fiscalist controversy down to this: Fiscalists believe a bond-financed increase in the budget deficit has *permanent* expansionary effects. Monetarists believe that an increased budget deficit, if unaccompanied by more rapid monetary expansion, will leave excess demand for goods virtually unchanged, because a rise in the financial wealth/money ratio will raise interest rates and "crowd out" private investment. For monetarists, bond-financed increases in deficit spending have *temporary* stimulating effects which fade away as lower private investment offsets higher government spending.[1]

Stein (1974) offers a general model in three differential equations which encompasses both fiscalist and monetarist views as special cases. His 1976 paper provides empirical estimates of *two* of these three equations in integral form:

$$(1)\ \begin{bmatrix} U \\ \pi \end{bmatrix} = R\begin{bmatrix} U_{-1} \\ \pi_{-1} \end{bmatrix} + S\begin{bmatrix} DG_{av} \\ \mu_{av} \\ \lambda_{av} \end{bmatrix} + \begin{bmatrix} R_{10} \\ R_{20} \end{bmatrix}$$

where U (unemployment rate) and π (rate of price change) are state variables; DG (dollar change in real government purchases per unit of capital), μ (growth rate of money stock), and γ (growth rate of public interest-bearing debt) are control variables, all averaged over the preceding three quarters; R and S are 2 x 2 and 2 x 3 matrices of coefficients; and R_{10} and R_{20} are constants of integration.

The model's cynosure and alleged focus of monetarist-fiscalist controversy are coefficients S_{13} and S_{23}, the partial derivatives of U and of π with respect to γ. If $S_{13} \geqslant 0$ and $S_{23} \leqslant 0$, an upward shift in the *IS* curve (without higher monetary growth) will be offset by an upward shift in the *LM* curve—the monetarist position. If $S_{13} < 0$ and $S_{23} > 0$, the *IS* curve shifts by more than the *LM* curve—the fiscalist position.[2] Stein's empirical estimates for 1960.IV-1973.IV and two subperiods are consistent with $S_{13} \geqslant 0$ and $S_{23} \leqslant 0$, the monetarist position.

In Stein (1974), the expected rate of price change π^* determines, along with excess demand for labor, the rate of growth of money

*Visiting lecturer, economics department and Woodrow Wilson School, Princeton University, and associate professor, Technion-Israel Institute of Technology. I gratefully acknowledge financial support from the Sloan Grant for Applied Microeconomic Research, Princeton University, and from the Foerder Institute for Economic Research, Tel Aviv University. I wish to thank George de Menil and Joseph A. Livingston for supplying me with their data, Haim Nisenson and Yael Benyamini for research assistance, and George Borts for spurring me to broaden the expectations data used in earlier drafts.

[1]Apparently John Maynard Keynes himself was not a fiscalist or neo-Keynesian. In a post-*General Theory* exchange with Dennis Robertson, he expressed full awareness of crowding out (which he termed "congestion in the capital markets"), and finally assented "... in general, the banks hold the key position in the transition from a lower to a higher scale of activity. If they refuse to relax, the growing congestion of the short-term loan market or of the new issue market, as the case may be, will inhibit the improvement..." (1937, p. 668). (See also passages in *The Collected Writings of John Maynard Keynes*, pp. 209, 229.) Richard Kahn also notes that "any attempt to increase employment may be rendered nugatory [by the inaction of the Central Bank]" (p.3). For this point and related passages, I am indebted to James Trevithick. For a modern neo-Keynesian view of crowding out, see James Tobin.

[2]In Stein (1974), the coefficient whose sign determines whether the model leads to monetarist or fiscalist results is P_5, defined as

$$P_5 = \frac{\partial \pi}{\partial E}\left[\frac{\partial C}{\partial \theta} + \frac{\partial \rho}{\partial \theta}\frac{\partial I}{\partial \rho}\right]$$

where ρ is the interest rate, π is the rate of price increase, E is excess demand for goods, C and I have their conventional meanings, and θ is the financial wealth/money ratio. The bracketed term is the impact of an increase in public debt/money on excess demand. It will be negative (and so will P_5) if the added consumption induced by wealth effects is less than the amount of private investment choked off by higher interest rates. The coefficient P_5 will be positive if the added consumption exceeds the fall in investment.

wages, and also influences excess demand for bonds. Thus, in reduced-form differential equations for DU and $D\pi$, both π and π^* appear as explanatory variables. In looking "Inside the Monetarist Black Box" Stein drops π^*, the third state variable and corresponding equation in his 1974 model, on the grounds that "π^* will be highly correlated with π" and "any method of deriving the expected rate of price change is arbitrary" (p. 209). Since inflation expectations figure importantly in the 1974 model, the absence of π^* from Stein's 1976 estimates leaves several interesting questions unanswered.

Two such questions are: To what extent are the empirically determined "monetarist" values of S_{13} and S_{23} influenced by omission of π^*? Is π^* determined by an underlying monetarist or fiscalist mechanism; that is, do people *believe* bond-financed deficit spending to be permanently expansionary? Taking up suggestions by Stein himself and Carl Christ,[3] this study exploits data on directly observed inflation expectations of both experts and laymen to restore π^* to Stein's model and reestimate it.

It is found that (a) addition of the expected rate of price change as an explanatory variable to regression equations in U and in π does not alter the monetarist conclusion that the partial derivative of excess demand for goods with respect to the ratio financial wealth/money is weakly negative; (b) the evidence tends to support the contention that the structure of inflation expectations is monetarist rather than fiscalist in nature for both experts and laymen.

I. Source and Nature of Inflation Expectations Data

Six different inflation expectations series were collected and screened for possible use. It was found that of the six, the first three listed below (*LIV*, *DMB*, *JUS*) contain virtually all of the information to be found in answers to survey questions upon which the six series were based. It is these three series which are used in the next section.

[3]"... the expected rate of price change remains elusive. I hope it will stimulate others to develop better estimates" (Stein, 1976, p 210). See also Christ (p. 238).

LIV is the expected rate of increase in the Consumer Price Index during the coming twelve months, based on a biannual survey of a panel of economists from government, banking industry and universities conducted by Joseph A. Livingston, nationally syndicated economic columnist. To produce quarterly estimates from twice-yearly surveys, simple interpolation was used.

DMB is from unpublished data appendix in George de Menil and Surjit Bhalla for 1966.II–1973.II, calculated directly from answers to the University of Michigan's Survey of Consumers question: "... would you say that a year from now prices will be about 1 or 2% higher, or 5%, or closer to 10% higher than now or what?"; prior to 1966.II, calculated indirectly from answers to the same survey's question: "... do you think (prices of the things you buy) will go up in the next year or go down, or stay where they are now?," essentially by establishing the relationship between answers to the two questions during the period when they overlapped and extrapolating that relationship backward.

JUS is from F. Thomas Juster (pp 28–30); utilizing only answers to an "up/down/stay the same" question, Juster assumes inflation expectations are normally distributed and solves the following two equations for two unknowns (x, mean inflation expectation, and s, its variance):

$$a = (1.25\% - x)/s$$
$$b = (-1.25\% - x)/s$$

where the interval $(-1.25\% + 1.25\%)$ is the amount of expected price change within which respondents answer "no change," and a and b are found from the percent answering "up" $= 1 - F(a)$ and the percent answering "down" $= F(b)$, where $F(\cdot)$ is the cumulative normal distribution.[4]

LIVH is the expected (annual) rate of

[4]There is evidence that the assumption of normally distributed inflation expectations may be only partially valid (see John Carlson). Further, Carlson and Harl Ryder note a number of serious difficulties with this method of extracting quantitative expectations from a qualitative question. For empirical studies of inflation expectations computed in this manner, see de Menil for the United States and Carlson and Michael Parkin for Britain.

increase in Consumer Price Index during the coming *six* months; also from Livingston survey.

JSDIR is from Juster. Somewhat similar to *DMB*, this series is based on the two questions quoted above, together with answers to a supplementary question: "Do you think that prices will be a lot higher (lower) or only a little higher (lower) next year?," asked in surveys during 1951.IV–1959.II.

DM is from de Menil, very similar to *JUS*.

For the period 1960.IV–1970.IV, zero order correlation coefficients between *LIV* and *LIVH* were 0.98; between *JUS* and *DM*, 0.96; and between *DMB* and *JSDIR*, 0.97. For this reason, only *LIV* was used below to reflect experts' inflation expectations, and *JUS* and *DMB* were used to measure popular inflation expectations, with *JUS* based on a "qualitative" survey question, and *DMB* based in part on a quantitative survey question. The series *LIV*, *JUS*, and *DMB* can be said to reflect fairly well the available data on directly observed expectations.

II. Empirical Results

Tables 1 and 2 present empirical estimates of equation (1), both excluding π^* as an explanatory variable and including it.[5]

Unemployment Rate U(t). The nonnegativity of S_{13}, the coefficient of γ, is maintained when π^* is added, for both 1960.IV–1970.IV and 1960.IV–1973.IV.

Coefficients of π^* itself are (with the possible exception of *JUS*) nonnegative, consistent with the posited inability of anticipated inflation to lower real wages and unemployment. The predicted inverse relation between unemployment and monetary growth, and between unemployment and changes in real government spending, is virtually unaffected by the presence of π^*.

Rate of Price Change $\pi(t)$. As with $U(t)$, the crucial coefficient of γ retains its monetarist sign (negative, and statistically significant at 0.10) when π^* is added to the

[5]All the data used in this section are available on request from the author

TABLE 1—THE UNEMPLOYMENT RATE EQUATION[a]

$U(t)$ regressed on		1960.IV–1970.IV				1960.IV–1973.IV			
Symbol	Variable	Stein (1976)	*LIV*	*DMB*	*JUS*	Stein (1976)	*LIV*	*DMB*	*JUS*
	Constant	5.5	5.6	5.5	5.7	4.1	4.3	4.3	4.5
		(6.0)	(6.8)	(6.0)	(6.8)	(5.8)	(6.3)	(6.0)	(6.5)
$U(t-3)$	Lagged Unemployment Rate	0.16	0.16	0.16	0.15	0.34	0.35	0.33	0.32
		(1.2)	(1.3)	(1.3)	(1.2)	(3.3)	(3.3)	(3.2)	(3.0)
$\pi(t-3)$	Lagged Inflation Rate	−0.06	−0.14	−0.08	−0.05	0.05	−0.02	0.03	0.03
		(−0.6)	(−0.97)	(−0.7)	(−0.5)	(0.6)	(−0.2)	(0.3)	(0.4)
$\pi^*(t-3)$	Lagged Expected Inflation Rate	-	0.10	0.06	−0.02	-	0.08	0.01	−0.03
			(0.71)	(0.37)	(−0.42)		(0.81)	(0.10)	(−0.8)
$DG(t-\delta)$	\$ Change in Real Government Purchases per Worker	−0.005	−0.006	−0.005	−0.005	−0.004	−0.005	−0.005	−0.005
		(−3.8)	(−4.6)	(−4.5)	(−4.3)	(−3.3)	(−4.0)	(−3.9)	(−4.0)
$\pi(t-\delta)$	Percent Change in Money Supply	−0.32	−0.34	−0.33	−0.33	−0.24	−0.27	−0.26	−0.26
		(−6.0)	(−7.2)	(−7.1)	(−7.0)	(−5.2)	(−6.2)	(−6.2)	(−6.1)
$\gamma(t-\delta)$	Percent Change in Public Debt	0.05	0.06	0.06	0.06	0.09	0.10	0.11	0.11
		(1.0)	(1.2)	(1.3)	(1.3)	(2.1)	(2.4)	(2.4)	(2.5)
Standard Error		0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
R^2		0.71	0.79	0.79	0.79	0.6	0.68	0.68	0.68
Durbin-Watson		1.2	1.2	1.2	1.2	1.0	1.1	1.0	1.1

Sources: U, π, DG, μ, γ, from Stein (1976); π^*, see text.

[a]To eliminate serial correlation, all variables were transformed by: $X-\rho X_{-1}$; t-statistics are shown in parentheses.

TABLE 2—THE RATE OF PRICE CHANGE EQUATION[a]

$\pi(t)$ regressed on: Symbol	Variable	1960.IV–1970.IV Stein (1976)	LIV	DMB	JUS	1960.IV–1973.IV Stein (1976)	LIV	DMB	JUS
	Constant	2.0	1.87	1.91	1.91	−1.0	2.9	2.3	2.6
		(2.3)	(2.5)	(2.1)	(2.3)	(−0.5)	(0.3)	(0.03)	(0.03)
$U(t-3)$	Lagged Unemployment Rate	−0.3	−0.3	−0.3	−0.3	−0.34	−0.36	−0.35	−0.36
		(−2.4)	(−2.4)	(−2.4)	(−2.4)	(−1.8)	(−1.7)	(−1.8)	(−1.8)
$\pi(t-3)$	Lagged Rate of Inflation	0.64	0.41	0.63	0.62	0.16	0.18	0.10	0.16
		(6.9)	(3.0)	(6.3)	(6.6)	(0.8)	(0.8)	(0.5)	(0.8)
$\pi^*(t-3)$	Lagged Expected Inflation Rate	-	0.29	0.03	0.02	-	0.09	0.32	0.12
			(2.2)	(0.20)	(0.4)		(0.6)	(2.1)	(1.8)
$DG(t-\delta)$	$ Change in Real Government Purchases per Worker	−0.0	−0.0	−0.0	−0.0	−0.003	−0.003	−0.003	−0.003
		(−0.3)	(−0.7)	(−0.4)	(−0.4)	(−1.2)	(−1.2)	(−1.4)	(−1.2)
$\mu(t-\delta)$	Percent Change in Money Supply	0.17	0.17	0.18	0.18	0.04	0.07	0.08	0.07
		(3.5)	(3.9)	(4.0)	(4.0)	(0.5)	(0.8)	(0.9)	(0.8)
$\gamma(t-\delta)$	Percent Change in Public Debt	−0.09	−0.11	−0.10	−0.10	−0.16	−0.14	−0.13	−0.13
		(−1.9)	(−2.5)	(−2.0)	(−2.0)	(−2.1)	(−1.8)	(−1.7)	(−1.7)
Standard Error		0.23	0.22	0.23	0.23	0.38	0.40	0.38	0.39
R^2		0.9	0.91	0.89	0.89	0.26	0.24	0.30	0.28
Durbin-Watson		1.6	2.0	1.7	1.8	1.5	1.3	1.4	1.5

[a]To eliminate serial correlation, all variables were transformed by: $X-\rho X_{-1}$, t-statistics are shown in parentheses.

regression for both 1960.IV–1970.IV and 1960.IV–1973.IV. The implication is that bond-financed deficits reduce private investment (by raising interest rates) by more than they increase private consumption (by increasing privately held wealth), even when inflation expectations are taken into account.

For the two periods, three of the six π^* coefficients are statistically significant and positive. A comparison of these coefficients indicates that lagged *expected* inflation had roughly similar (in one case, substantially greater) impact on current inflation, compared with lagged *actual* inflation.

The strongly significant coefficient of μ, the money supply control variable, for 1960.IV–1970.IV becomes small in magnitude and statistically insignificant for 1960.IV–1973.IV, with expectations present or absent. Wage and price controls may provide a partial explanation. In general, adding the 1970's data to regression estimates alters most coefficients substantially and is indicative of major structural change, with traditional relationships among economic variables seriously fractured.

Expected Rate of Price Change $\pi^*(t)$. Table 3 presents multiple regression estimates for each of the three inflation expectations series as dependent variable. Across the three series, there are some important similarities

The coefficients of lagged unemployment and lagged inflation are negative and positive, respectively, as theory predicts, and for all but two of the twelve coefficients, equal or exceed their standard deviations. For 1960.IV–1970.IV, coefficients of lagged expected inflation are not statistically significant, but become significant at 0.10 for 1960.IV–1973.IV; this suggests that expectations became a more important part of the inflationary process from about 1970 on. Control variables DG and μ are significant and have correct signs for 1960.IV–1970.IV, but (with one exception) lose statistical significance when the three years to 1973.IV are added. Did the wage-price controls of these years lead both experts and laymen to replace fiscal and monetary policy variables with other information in forming their inflation expectations?[6] If so, with what other information?

[6]For all three expectations series, inclusion of 1971–73 increases the standard error of estimate by one-half or more.

TABLE 3—THE EXPECTED RATE OF PRICE CHANGE EQUATION

$\pi^*(t)$ regressed on: Symbol	Variable	1960.IV–1970.IV LIV[a]	DMB[b]	JUS[a]	1960.IV–1973.IV LIV[a]	DMB[a]	JUS[b]
	Constant	0.37	2.4	4.0	2.01	3.2	4.8
		(0.4)	(2.8)	(3.2)	(1.3)	(1.9)	(2.1)
$U(t-3)$	Lagged Unemployment	−0.08	−0.11	−0.4	−0.22	−0.31	−0.6
	Rate	(−0.6)	(−1 2)	(−2.3)	(−1.0)	(−1.4)	(−1.8)
$\pi(t-3)$	Lagged Inflation	0.57	0 12	0.38	0.22	−0.20	−0.14
	Rate	(3 2)	(1.7)	(2.6)	(1.0)	(−1.1)	(−0.6)
$\pi^*(t-3)$	Lagged Expected	0.18	0.11	−0.13	0.32	0.60	0.34
	Rate of Inflation	(1.0)	(0.6)	(−0.8)	(1.6)	(2.5)	(1.7)
$DG(t-\delta)$	Change in Real	0.004	0.001	0.004	−0.001	−0.002	−0.00
	Government Purchases	(3 1)	(1.5)	(3.4)	(−0 2)	(−0.9)	(−0.1)
$\mu(t-\delta)$	Percent Change	0.14	0.09	0.108	0.16	0.05	0.12
	in Money Supply	(2.8)	(2 9)	(1.7)	(1 6)	(0.5)	(1.0)
$\gamma(t-\delta)$	Percent Change	−0.16	−0.05	−0.10	−0.04	0.07	0.03
	in Public Debt	(−3.0)	(−1.6)	(−1.6)	(−0.4)	(0.7)	(0.2)
Standard Error		0.30	0.25	0.57	0 48	0.52	0.90
R^2		0 38	0.81	0.58	0 18	0.10	0 26
Durbin-Watson		1.9	2 0	2.2	1 4	1 6	1.7

[a]Transformed by $X - \rho X_{-1}$ to eliminate serial correlation; t-statistics are shown in parentheses.
[b]Not transformed (no serial correlation present)

For all three expectations series, the coefficient of γ is negative or zero (although significantly only for *LIV*, 1960.IV–1970.IV, and approaching significance for *DMB*). This supports the monetarist view of expected inflation, which contends that both experts and laymen do not *perceive* larger bond-financed deficits to have permanent expansionary effects.

There is another sense in which inflation expectations reveal their monetarist structure. Let us set aside the issue of the statistical significance of control variables' coefficients and look at the *magnitude* of those coefficients. To facilitate comparison, *beta* coefficients for DG, μ, and γ were calculated and tabulated in Table 4, for U, π and π^* as dependent variables. (*Beta* coefficients standardize for different units of measurement by multiplying regression coefficients by s_x/s_y, where s_x and s_y are the standard deviations of the explanatory and dependent variables, respectively.)

A unit change in monetary growth has the largest impact on inflation expectations for four out of the six cases (three expectations series and two periods), though in one case the difference is very small. For 1960.IV–1970.IV, between nearly half to more than all of a unit change in DG is offset by a unit

TABLE 4—*Beta* COEFFICIENTS OF EXPLANATORY POLICY VARIABLES (DG, μ, γ)[a]

	Dependent Variable	1960.IV–1970.IV DG	μ	γ	1960.IV–1973.IV DG	μ	γ
$U(t)$	Unemployment	−0.39	−0.67	0.10	−0.38	−0 66	0.27
$\pi(t)$	Rate of Price Change	−0.02	0.27	−0.11	−0.15	0.13	−0.23
$\pi^*(t)$	Expected Rate of Price Change						
	LIV	0.21	0.23	−0.19	0.05	0.02	−0.18
	DMB	0.15	0.38	−0.16	−0.17	0.27	0.15
	JUS	0 31	0.21	−0.15	0.01	0.30	0.03

[a]*Beta* coefficients are ordinary regression coefficients multiplied by s_x/s_y, where s_x and s_y are the standard deviations of the independent and dependent variables, respectively.

change in γ, in terms of its impact on expected inflation. This suggests that even *temporary* expansionary effects of *DG*—when unaccompanied by higher μ—are perceived as small.

For unemployment, monetary growth's impact is more than 70 percent larger than changes in real government spending. For the actual rate of price change, μ appears most powerful for 1960.IV–1970.IV but least powerful for 1960.IV–1973.IV.

One curious result shown in Table 4 is that for *popular* inflation expectations, monetary growth has a much larger (for *DMB*) or about equal (for *JUS*) impact on the expected rate of price change among laymen than it does on experts' expectations (*LIV*). One wonders precisely how this works. It is unlikely that pipefitters and stevedores furrow their brows over the latest figures on M_1, M_2, and excess reserves. Perhaps people are sensitive to changes in interest rates and installment credit. But George Katona has shown us that "most American consumers are not well informed about how much installment credit costs. (Surveys show) only 20 percent, or at most 30 percent indicated . . . some awareness of actual interest rates" (p. 279). Further theoretical and empirical work at the micro level on the structure of expectations is justified.

III. Conclusion

Restoring inflation expectations to Stein's blueprint of the monetarist black box has left his empirical monetarist findings intact. Further, the *perceptions* of both experts and oridinary people appear to be monetarist. To some extent, those *perceptions* themselves are a black box whose mechanism need careful study, all the more so since expectations are likely to remain an important element in the inflationary process for some time to come.

REFERENCES

J. A. Carlson, "Are Price Expectations Normally Distributed," *J. Amer. Statist. Assn.*, Dec. 1975, *70*, 749–54.

——— **and M. Parkin,** "Inflation Expectations," *Economica*, May 1975, *42*, 123–38.

——— **and H. Ryder,** "Quantitative Expectations from Qualitative Surveys: A Maximum-Likelihood Approach," mimeo., Purdue Univ. and Brown Univ., Oct. 1973.

C. Christ, "Comments on Stein," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976

G. de Menil, "The Rationality of Popular Price Expectations," mimeo., Princeton Univ. and INSEE, Paris, Dec. 1974.

——— **and Surjit S. Bhalla,** "Direct Measurement of Popular Price Expectations," *Amer. Econ. Rev.*, Mar. 1975, *65*, 169–80.

Elizabeth Johnson, *The Collected Writings of John Maynard Keynes*, Vol. 14, New York 1971.

F. T. Juster, "Uncertainty, Price Expectations and the Personal Saving Rate," in Burkhard Strumpel et al., eds., *Surveys of Consumers 1972–73*, Inst. Soc. Res., Univ. of Michigan 1975.

Richard Kahn, *Selected Essays on Employment and Growth*, Cambridge 1972.

George Katona, *Psychological Economics*, Elsevier 1975.

J. M. Keynes, "The 'Ex-Ante' Theory of the Rate of Interest," *Econ. J.*, Dec. 1937, *47*, 663–69.

J. A. Livingston, surveys published twice yearly, *Philadelphia Inquirer*, 1948 on.

J. Stein, "Unemployment, Inflation and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867–87.

———, "Inside the Monetarist Black Box," in his *Monetarism*, Amsterdam 1976.

J. Tobin, "Deficit Spending and Crowding Out in Shorter and Longer Runs," disc. paper no. 463, Cowles Foundation, Yale Univ., Sept. 1977.

The Wealth-Age Relation among the Aged

By THAD W. MIRER*

Aged people face the problem of allocating over the rest of their lifetimes the wealth that they have accumulated during their working years. This problem is fairly complex, because they are uncertain about the dates of their death and because they may wish to use their wealth to finance general consumption, or hold it to meet emergencies, leave estates, or maintain power or status. Whether skilled at dynamic programming or not, aged people effectively do make these decisions. What are the results?

Some studies of the entire lifetime pattern of saving and wealth tend to support the simple life cycle model, in which the aged use up their wealth to help finance consumption. Harold Lydall, for example, found saving rates to increase and then decrease with age in the whole population and to be negative for the group aged 65 and over. Guarded support for the theory was found by Franco Modigliani and Albert Ando in Britain and by Dorothy Projector in the United States. More recently, A.F. Shorrocks examined the lifetime pattern of wealth holding in estate-tax data for a single cohort and found the hump pattern that is characteristic of life cycle saving.

Other evidence on wealth holding runs contrary to this, however. In a study of estate-tax returns in Washington, D.C., James D. Smith found that wealth increased with age among the aged, after controlling for other demographic variables. Examining national estate-tax data, John Brittain found wealth to increase with age in the United States and A. B. Atkinson and A. J. Harrison reported the same pattern in England and Wales. In a sample of (*TIAA*) annuity recipients analyzed by James Mulanaphy, over half reported that their total savings had increased during retirement and another quarter reported that their savings had remained about the same.

In this paper, I examine wealth holding patterns among aged married couples, using survey data representing the entire aged population of the United States in 1968. The statistical analysis serves as a test of the hypothesis that emerges from life cycle theory and provides some further exploration into the economic behavior of the aged. I find that the aged tend to increase their wealth over time.

I. Life Cycle Theory and the Wealth of the Aged

The life cycle theory of saving,[1] which is usually applied to explain behavior over a whole economic lifetime, provides a framework for analyzing the wealth holding behavior of the aged. In its simplest form, which has had considerable success in explaining aggregate saving behavior (see Ando and Modigliani, and James Tobin), the theory assumes that intertemporal consumption smoothing is the only motive for holding wealth. Normally, people would accumulate wealth when they are young and working, and decumulate it when they are old and retired. Thus, the theory can be tested by examining the wealth-age relation: if wealth does not decrease with age among the aged, the theory is contradicted.

A formal model of the wealth holding behavior of the aged, based on the seminal work of Menahem Yaari, can be used to develop more specific hypotheses about the shape of the wealth-age profile, but it is not presented here. Yaari examined optimal consumption planning for an individual with an uncertain lifetime—a feature which is surely important in any model hoping to

*State University of New York-Albany. I wish to thank Wayne Finegar of the Social Security Administration for help in obtaining the data, and George Borts and a referee for helpful comments.

[1]Modigliani and Richard Brumberg might be considered the modern fathers of life cycle saving theory, but Roy Harrod had earlier shown that life cycle saving implies a hump-shaped pattern of wealth holding over the entire lifetime.

describe the behavior of the aged. In relevant cases, the exact shape of the resulting wealth-age profile is ambiguous, but the test proposed above remains generally appropriate.

A broader form of life cycle theory, which would include other motives for holding wealth—for bequests, for emergencies, and for power or status—could accommodate many of the criticisms which are sometimes made regarding the theory. However, such a broad form would be so inclusive that it would be difficult to test among the aged. No particular wealth-age relation would contradict the theory, for example.

Of these additional motives for holding wealth, the desire to leave a bequest is the only one to have received much attention in the literature. Some evidence weighs against attaching too much importance to the bequest motive, however. In the 1962 Survey of Financial Characteristics of Consumers (see Dorothy Projector and Gertrude Weiss, Table A30) only 4 percent of the respondents aged 65 and over cited "providing an estate" or a similar goal as a saving objective; 47 percent cited "providing for old age" and 34 percent cited "providing for emergencies." Carl Shoup (p. 15) notes that a "surprisingly small" amount of *inter vivos* giving takes place in view of the substantial bargain that exists in transferring wealth under gift taxes rather than under estate taxes. Also, it should be noted that the existence of bequests is not evidence of a strong bequest motive, since many people will die with positive wealth if death is uncertain. The precautionary and power motives probably deserve more attention, but it would seem to be quite a difficult task to disentangle them from the effects of a bequest motive until much richer data, and theory, become available.

II. Empirical Analysis

The cross-section survey data examined here will be used to construct wealth-age profiles and test the simple formulation of the life cycle hypothesis presented above. Unfortunately, the data are not so rich as to allow us to estimate a complete model of wealth holding or even to relate current wealth to wealth at retirement or lifetime income. The data are discussed in Part A. In Part B, the wealth-age relation is analyzed in the cross section, and in Part C the data are adjusted so that inferences regarding the lifetime behavior of individuals can be made. In Part D I briefly explore the composition of wealth by age.

The analysis is restricted to aged married couples in order to strengthen the longitudinal inferences that can be made. A cross section of aged couples may be thought of as the survivors of successive cohorts of married couples who have reached age 65. When intercohort differences are controlled for, the cross section is a relatively homogeneous sample of economic units from which we may try to infer lifetime behavior.[2] It may well be that widows, widowers, and never-married persons all behave quite differently from married couples, and therefore the findings here should not be extrapolated to them.

A. *The Data*

The data examined are from the 1968 survey of the Demographic and Economic Characteristics of the Aged (*DECA*), which provides employment, income, Social Security, and wealth information for 1967 on the aged population in the United States. This survey, which was conducted by the Census Bureau for the Social Security Administration, contains completed interviews with a representative sample of 8,248 aged persons. A complete description of the survey and some of its findings are presented by Lenore Bixby et al.

The *DECA* survey asked a number of questions about assets and debts which have been combined to yield estimates of the value of financial assets and of home equity for each couple. Financial assets include holdings of money in financial institutions plus the value of holdings in bonds, stocks, mutual funds, and personal loans and mortgages. Home equity is based on the respondent's estimate of current market value and the outstanding

[2]Substantial variation of death, divorce, or marriage rates by the level of wealth would weaken this statement. The impact of death rates is discussed below.

mortgage and other debt on the property. The sum of financial assets and home equity is taken here as total wealth, although it is not literally this because it excludes certain items (such as business or farm assets and equity in rental property) about which the survey did not ask specific questions.[3] Also, the capital values of pensions, annuities, Social Security, and life interests in estates are not considered.[4] In Parts B and C below I focus only on total wealth, and in Part D I examine its composition.

In general, wealth data derived from interview surveys are inferior to those for some other economic variables because of the greater difficulty of obtaining correct estimates. Wealth items tend to be underestimated,[5] and in *DECA* it appears that financial assets are underreported to a greater degree than home equity. No adjustments are made here to correct for these deficiencies. Couples for which any wealth item was ambiguous were deleted.

The age of a couple is taken to be the age of the husband. The actual age was determined from Social Security records matched with the interview record, rather than from the interview itself. Couples whose age was less than 65 were deleted from further study.

The only measure of preretirement circumstances is the level of education, classified here in four groups: 0–6 years; 7–8 years; some high school; some college. Couples whose education was not clearly reported were deleted. The education variable is used in some of the analysis below because it serves to divide the sample into subsamples that are more homogeneous with respect to wealth at retirement, length of life, and preferences regarding wealth holding than is the entire sample.

The basic sample used here consists of 2,713 observations. The couples range in age 65–99, with 37 percent being 65–69, 30 percent being 70–74, 21 percent being 75–89, 10 percent being 80–84, and 3.2 percent being 85 or over.[6] The youngest men were born in 1902 and entered the work force about the time of World War I. By today's standards, these couples are not highly educated: 30 percent had 0–6 years of schooling, 34 percent had 7–8, 29 percent had some high school education, and 10 percent had some college. The mean total wealth (in 1967 dollars) is over $19 thousand. The median is less, at $12 thousand, and the lower and upper quartiles (i.e., the twenty-fifth and seventy-fifth percentiles) are about $4 thousand and $24 thousand, respectively. Overall, total wealth is split roughly half and half between financial assets and home equity. While not millionaires, most aged couples have enough wealth to lead them to think seriously about what they should do with it.

B. *The Cross Section*

The cross-section patterns of wealth holding among the aged are examined in this section by regression techniques and by calculating percentile values from grouped data. For convenience in discussing the results, I shall use some terms that would be appropriate if the cross section were being interpreted as the lifetime behavior of individuals.

Using regression techniques to describe the age pattern of any variable in cross-section data necessitates a choice of functional form. Graphical and tabular analysis indicated that a simple regression of the form

$$(1) \qquad WEALTH = \alpha + \beta \cdot AGE$$

might suffice. This regression was estimated

[3]The excluded items constitute about 28 percent of the total wealth of aged consumer units as reported in the 1962 Survey of Financial Characteristics of Consumers (see Projector and Weiss, Tables A8 and A10); by contrast, the excluded items constitute about 45 percent of the total wealth of consumer units whose head was age 35–54.

[4]To some extent, perhaps a great one for many people, pension and Social Security programs tend to institutionalize the tenets of life cycle theory. Whether people would follow the same saving behavior in the absence of these institutions is an interesting, and open, question. In this paper, no further attempt is made to treat this form of "wealth."

[5]Robert Ferber has shown that misreporting of assets tends to be related to the level of wealth.

[6]The small number of observations with high ages opens the possibility for a few outliers in the data to seriously distort the wealth-age patterns. As a check, much of the analysis was repeated for couples age 65–84 only. the wealth-age patterns are very similar.

TABLE 1—WEALTH-AGE REGRESSIONS

		Constant (α)	Grades 7–8	High School	College	*AGE* (β)	R^2
Unadjusted Data	(1)	20,446	-	-	-	−141.7	.001
		(906)				(87.8)	
	(2)[a]	10,914	6,055	12,477	25,666	+20.82	.081
		(1,244)	(1,235)	(1,239)	(1,789)	(85.45)	
Adjusted Data	(3)	20,202	-	-	-	+306.6	.003
		(1,087)				(105.4)	
	(4)[a]	8,751	7,303	15,037	30,580	+502.0	.083
		(1,494)	(1,482)	(1,553)	(2,147)	(102.6)	

Note: Standard errors are shown in parentheses.
[a]In regressions (2) and (4), the coefficients for the three education groups measure the increment for that group over the excluded group (grades 0–6), whose wealth at retirement is given by the constant.

using data for the entire sample as well as for each education level subsample separately. Since the regressor *AGE* was actually defined as years of age over 64, α estimates wealth on the eve of retirement (i.e., at age 64) and β measures the annual change in wealth.

The results for the entire samples are presented in row (1) of Table 1. Mean wealth on the eve of retirement (α) is estimated to be $20,446 and the annual decrease in wealth (β) is estimated to be $141.7. This estimated coefficient for the change in wealth just barely misses being significant at the .10 level. The economic importance of this decrease may be gauged by computing the percentage change in wealth (relative to wealth on the eve of retirement) that occurs over the length of time equal to the expectation of life at age 65, which is about fifteen years. In this case, the fifteen-year change is −10.4 percent, which seems small relative to what one might anticipate if the aged were running down their assets to finance consumption. The extremely low R^2 statistic also tells us that wealth does not vary systematically with age very much at all.

An obvious alternative functional form for estimating the wealth-age pattern is the quadratic, which has often been used for estimating age-earnings profiles. In addition to allowing a simple test for nonlinearity, the quadratic is suggested by the simple life cycle theory in some cases. Accordingly, an equation including a squared *AGE* regressor was estimated, but this term was not significant. Hence, the relation may be considered essentially linear.

The regression analysis describes how average wealth varies with age in the entire sample. Could it be that those of modest means are forced to use up their wealth to finance consumption while the very wealthy are free to accumulate even more—thus leading to only a modest decline of wealth overall? To examine this question, the cumulative (weighted) relative frequency distribution of total wealth was computed for each of four five-year age intervals: 65–69, 70–74, 75–79, and 80–84; couples older than 84 were not grouped because the data are very thin in this range. The wealth values at various percentile positions in each distribution were calculated, and to facilitate comparisons a wealth-age "regression" was fit to each set of four linked percentile-age "observations" (for example, the ninety-fifth percentile value of wealth for ages 65–69, 70–74, 75–79, and 80–84). From these regressions, the fifteen-year percentage changes in wealth were computed.

The data show the lowest rates of decline in wealth (about 5 percent) at the first quartile and quintile points (below these levels, wealth holdings are negligible). The rate of decline increases with the percentile position in the distribution, being about 14 percent at the median and reaching 20 percent for couples at the eightieth through ninetieth percentiles. The rate of decline at the ninety-fifth percentile is about 13 percent. This pattern suggests a greater willingness to use up wealth the

greater is the initial level of wealth—except at the very highest levels.[7]

Thus, among the entire sample of the aged, the conclusion that wealth decreases with age in the cross section seems to hold broadly through the wealth distribution. At this level, the data are in rough qualitative agreement with the standard life cycle model. However, the decreases in wealth are quite modest compared to what might have been anticipated.

When we control for differences in education level even this modest pattern of decline vanishes. The education level groups are more homogenous within themselves than is the entire sample with respect to wealth at retirement and other factors (such as longevity and utility functions) that would determine patterns of wealth holding. Separate regressions based on equation (1) for each group allow estimation of different means for wealth at retirement and different patterns of wealth holding by age. In all cases, the estimated coefficient on age (β) was significant and the calculated fifteen-year changes in wealth ranged from -10 to $+20$ percent. In no case was an added squared *AGE* term significant.

An equation constraining the estimated β to be identical for all education groups while allowing for differences in wealth at retirement (α) was also estimated, with results shown in row (2) of Table 1. (A squared *AGE* regressor was insignificant.) An F-test accepted this constraining hypothesis. Based on that equation, wealth on the eve of retirement was $10,914 for couples with 0–6 years of schooling, $16,969 for those with 7–8 years, $23,391 for those with some high school education, and $36,580 for those with some college. The common yearly change of wealth showed an increase of $21, which is both insignificant and unimportant.

The fact that wealth barely changes with age when education is controlled for, while it decreases with age in aggregate, can be explained by observing the changes in education composition by age level. Because the population has become better educated over time, the proportion of the aged in the lowest education category increases with age while that in the highest category decreases. Since wealth at retirement increases with the level of education, we can see that one reason why older couples in the entire sample have less wealth than younger couples (i.e., why $\beta<0$) is that they started out at age 65 with less wealth.

C. *Correcting for Intercohort Differences in Wealth at Retirement*

Our interest is primarily in determining the wealth holding profiles of couples over their aged lifetimes rather than among different couples of various ages. The ideal data for this is longitudinal (panel) data from retirement to death on a sample of couples.[8] Lacking that, we can try to infer longitudinal behavior from the *DECA* cross-section data, but such inference is difficult. Basically we need to know and correct for all the differences that exist among cohorts.

For our purposes, probably the major difference among the cohorts in the cross section arises from the fact that successive cohorts have lived through times of increasing labor productivity and earnings occurring in the economy as a whole. Thus, successive cohorts have held greater and greater wealth at the time of retirement. This may be expected to hold within each education group as well as overall.

Assuming that a fairly strict set of assumptions actually hold, the simple adjustment we shall make to take into account the growth in wealth at retirement transforms the cross-section data from 1967 into the equivalent of (future) longitudinal data on couples who were on the eve of retirement in 1967. The key assumptions are 1) that the relative

[7]Shorrocks also found the decline to be less at the ninety-fifth percentile than at the ninetieth, but was not able to examine patterns among the less wealthy.

[8]The Social Security Administration is nearing completion of a ten-year longitudinal survey that will give data on a large sample of persons aged 68–73 at completion and will be valuable for examining the behavior of the aged in their "early" years. Lola Irelan provides an introduction to this Retirement History Survey.

distribution of wealth at retirement within each cohort was the same among cohorts, 2) that wealth at retirement has grown at the constant rate g from cohort to cohort, and 3) that because of constancy of behavior, the postretirement wealth-age profiles (relative to wealth at retirement) remain the same over time. If these assumptions do not hold, then our adjustment stands simply as a reasonable correction to the cross-section data that makes it more like longitudinal data.

If wealth at retirement grows at the rate g, then the typical profile for any given cohort is $(1 + g)$ times as high as that for the cohort that is one year older (i.e., that retired one year earlier) and it is $(1 + g)^t$ times as high as that for the cohort that is t years older. Thus the typical profile that will be experienced by the cohort of age 64 in our cross-section data will be $(1 + g)^t$ times as high as the profile for the cohort aged $64 + t$ in our data. Or, put differently, the wealth holding t years from now of the typical couple now age 64 will be $(1 + g)^t$ times as much as the current wealth holding of the typical couple now age $64 + t$. Thus if we inflate all observed wealth values in the cross section by the factor $(1 + g)^{AGE}$, then the wealth-age pattern in the adjusted data will trace out the future wealth holding profile of the typical couple now on the eve of retirement (i.e., age 64 in the data). This is the adjustment I make. (Note that $AGE =$ Calendar Age $- 64$.)

This procedure is illustrated in Figure 1 for the case in which wealth increases with age within each cohort. Because wealth at retirement is continually increasing, the actual wealth-age profiles fan out and upward to the right.[9] However, because the differences in wealth at retirement among cohorts are greater than the tendency to accumulate, the observed cross-section relation shows wealth decreasing with age. The adjustment procedure raises each observed point vertically in the figure, through multiplying by $(1 + g)^{AGE}$, so that the adjusted wealth values trace out the wealth-age profile that will

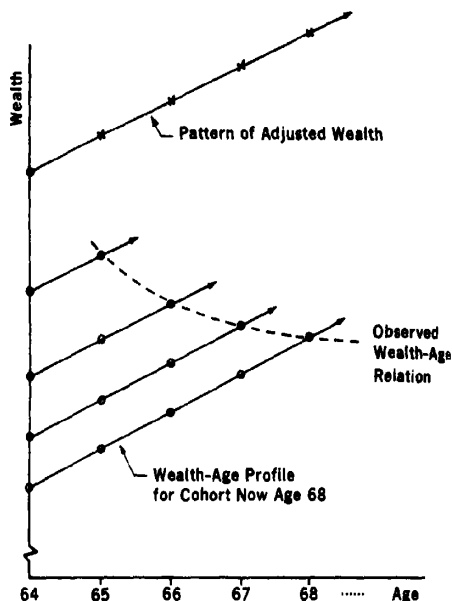



FIGURE 1

occur for the cohort now aged 64. If, in reality, each cohort's wealth-age profile is downward sloping, the procedure works similarly. The pattern of adjusted wealth observations will be flatter than the observed cross-section relation.

The wealth-age pattern in the adjusted data depends critically on the chosen value for the correction parameter. In addition to choosing a value that is conservative, in terms of its tendency to disprove the life cycle hypothesis, some sensitivity analysis will be presented.

Although we have no direct method for estimating the rate of growth of wealth at retirement, under commonly accepted assumptions in life cycle theory this rate is equal to the rate of increase in incomes between successive cohorts. Using unpublished data for 1937–72, Martin Feldstein calculated the rate of growth of real per capita disposable income to be about 2 percent per year, and we use this as our estimate of g. This is probably a conservative estimate, as real family incomes grew at about 3 percent in the relevant post-World War II period. My adjustment procedure leads to increasing the

[9]The profiles would not be parallel under the set of assumptions made above, but are drawn that way for convenience because the vertical scale is broken.

reported wealth values of couples aged 70 and 80, for example, by the factors 1.13 and 1.37, respectively.

Parallel to the regression analysis reported above, equations of the form

$$\text{(2)} \quad ADJUSTED\ WEALTH = \alpha + \beta \cdot AGE$$

were estimated for the entire sample and each education group. Overall, wealth is found to increase significantly with age (see regression (3) in Table 1). Over the fifteen-year period (corresponding to the expectation of life) after age 64, total wealth increases by about 23 percent in the entire sample. Again, an added squared *AGE* regressor was not significant.

Among the education groups, the increases of wealth with age were all fairly significant, and ranged 22–72 percent over a fifteen-year period. A regression with a common β (age effect) was also estimated (see row (4), Table 1), and an F-test indicates that we should accept the hypothesis of a common β. This equation shows a sizeable and significant tendency for wealth to increase during retirement.

How sensitive are the results to the chosen value of g? Regressions based on equation (2), analogous to regression (3) in Table 1, were estimated assuming 1 percent and 3 percent as alternative rates of growth. Compared to the fifteen-year increase of 23 percent found with the 2 percent growth adjustment to the data, the fifteen-year increase was 4 percent with the 1 percent adjustment and 46 percent with the 3 percent adjustment. Thus the wealth-age patterns reported here are fairly sensitive to the chosen rate.

Two additional caveats should be noted regarding these wealth-age patterns. First, another important difficulty of inferring longitudinal wealth patterns from cross-section data arises from the fact that there normally is a positive relation between economic status and longevity. Poor people tend to die off faster than rich people, so old people tend to look relatively rich. This imparts a positive bias to the slope of our wealth-age patterns. However, education is also highly correlated with longevity, and the results above show that when we separate the sample into education level groups, the rate of increase of wealth with age tends to be even greater than indicated in the aggregate. While differential mortality should not be simply ignored out-of-hand, it may be no more important than a host of other differences between cohorts which have not been accounted for here.[10]

Second, the sample data include couples who have received earnings from employment since age 64 as well as those who have not. While the data do have information on current employment status, it is not possible to isolate a sample who were completely retired from age 65 on. When aged people are employed, we would expect a greater tendency to increase wealth. Employment decreases with age among the aged, but we found no tendency for wealth to decrease even among the very old. As a further check, the basic adjusted wealth-age patterns were reestimated for sample members aged 75 and up and they too showed that wealth increases with age. The coefficient on *AGE* was moderately significant, and the increase amounted to about 23 percent (again!) over a ten-year period (equal to the expectation of life) after age 74.

D. *Changes in the Composition of Wealth*

What happens to the composition of wealth holdings as people age? Conventional wisdom holds that as retired people age their assets become increasingly concentrated in the equity of their home, and this belief seems to be one of the foundations for recent laws providing "circuit-breaker" property tax relief for the aged. The evidence above suggests that this may not be so, because the aged tend not to use up wealth at all.

This question can be explored further by examining the composition of wealth among aged homeowners in our sample.[11] For each

[10] In Shorrocks' work with estate-tax data, his correction for differential mortality was the crucial step that yielded a decreasing wealth-age relation.

[11] The results for the entire sample are similar.

aged couple, the ratio of home equity to total wealth (expressed in percentage points) is regressed on the couple's age and its total wealth,

$$\frac{HOME\ EQUITY}{WEALTH} = a + b \cdot AGE + c \cdot WEALTH \tag{3}$$

The hypothesis noted above suggests that $b > 0$, and we anticipate that $c < 0$, because housing provides a service necessary to homeowners. The regression was estimated for the total sample of homeowners and for the four education subsamples, and the results are reported in Table 2. (F-tests indicated that neither the c's nor the a's could be restricted to be equal across education groups. Hence the separate regressions are reported here.)

The age effect is consistently negative but insignificant in most cases, and it is never very important. (When the age effect is constrained to be equal for all education groups, it is negative and significant.) There is no evidence supporting the conventional hypothesis.

The scale effect, measured by the regression coefficient on total wealth, is negative and highly significant in all cases. The parameter estimates showing the percentage point decrease in the home equity/wealth ratio that occurs for every $1,000 of increased wealth range between −.36 and −.76. For the entire sample the estimate is −.52, indicating that a 70-year old couple with $50,000 of total wealth will have a home equity/wealth ratio of .59, which is about 10 percentage points lower than for a couple with $30,000 of total wealth.

TABLE 2—HOME EQUITY/WEALTH RATIO REGRESSIONS

Sample	Constant (a)	*AGE* (b)	*WEALTH* (c)	R^2
All	85.99	−.1323	− 5217	.290
(N = 2003)	(1.00)	(.0883)	(.0183)	
Grades 0–6	94.27	−.2565	−.7644	.262
(N = 523)	(1.86)	(.1521)	(.0573)	
Grades 7–8	85.49	−.0586	− 5486	.225
(N = 695)	(1.75)	(.1506)	(.0388)	
High School	84.38	− 2698	−.4874	.315
(N = 580)	(1.79)	(1802)	(.0302)	
College	77.42	−.2749	−.3636	.280
(N = 205)	(3.33)	(.2953)	(.0414)	

Notes: Parentheses contain standard errors. Wealth is measured in thousands of dollars, here only.

III. Conclusions

In view of the wide acceptance enjoyed by the life cycle theory of saving, the main empirical findings reported here are quite surprising. The raw cross-section data show that wealth (not considering the capital value of pensions, Social Security, etc.) declines modestly, or perhaps not at all, with age among the aged. An examination of the wealth-age profiles at various percentile positions within the distribution of wealth shows that this conclusion holds broadly at all levels of wealth. When adjustments are made to correct for intercohort differences in wealth at retirement, it is found that wealth clearly increases with age. While the adjustments do not transform the data into true longitudinal observations, it is hard to escape the conclusion that the aged do not run down their wealth during their lifetimes.

The major implication for economic theory of this finding is that the simple form of the life cycle theory of saving, in which wealth is accumulated during working years in order to finance consumption during retirement, is too simple. Precautionary, bequest, or other motives must be taken into account if the theory is to explain the wealth holding behavior of persons toward the end of their lives. This reformulation may be important for the saving-retirement-Social Security nexus in particular.

This finding also raises questions for public policy regarding the aged. How should their welfare be measured? Marilyn Moon used the simple life cycle model normatively to help construct a comprehensive measure of economic welfare, and Juanita Kreps used it to estimate consumption possibilities after retirement. But, if the normative model is at variance with positive behavior, these approaches may be invalid. A related question concerns how the benefits of income support programs should be related to the wealth of the aged: should there be mandatory spend downs, or a "tax" on wealth?

REFERENCES

A. Ando and F. Modigliani, "The Life-Cycle Hypothesis of Savings," *Amer. Econ. Rev.*, Mar. 1963, *53*, 55–84.

A. B. Atkinson and A. J. Harrison, *Distribution of Personal Wealth in Britain*, Cambridge 1978.

Lenore Bixby et al., *Demographic and Economic Characteristics of the Aged: 1968 Social Security Survey*, Washington 1975.

John Brittain, *Inheritance and the Inequality of Material Wealth*, Washington 1978.

M. Feldstein, "Social Security, Induced Retirement, and Aggregate Capital Accumulation," *J. Polit. Econ.*, Sept./Oct. 1974, *82*, 905–26.

R. Ferber et al., "Validation of a National Survey of Consumer Financial Characteristics: Savings Accounts," *Rev. Econ. Statist.*, Nov. 1969, *51*, 436–44.

Roy Harrod, *Toward a Dynamic Economics*, London 1948.

L. M. Irelan, "Retirement History Study: Introduction," *Soc. Sec. Bull.*, Nov. 1972, *35*, 3–8.

Juanita Kreps, *Lifetime Allocation of Work and Income*, Durham 1971.

H. Lydall, "The Life Cycle in Income, Saving, and Asset Ownership," *Econometrica*, Apr. 1955, *23*, 131–50.

F. Modigliani and A. Ando, "Tests of the Life Cycle Hypothesis of Savings," *Bull. Oxford Univ. Inst. Econ. Statist.*, May 1957, *19*, 99–124.

——— **and R. Brumberg**, "Utility Analysis and the Consumption Function: An Interpretation of Cross-Section Data," in Kenneth Kurihara, ed., *Post-Keynesian Economics*, New Brunswick 1954.

M. L. Moon, "The Economic Welfare of The Aged and Income Security Programs," *Rev. Income Wealth*, Sept. 1976, *22*, 253–69.

James Mulanaphy, *Statistical Report: 1972–73 Survey of Retired TIAA-CREF Annuitants*, New York 1974.

Dorothy Projector, *Survey of Changes in Family Finances*, Washington 1968.

——— **and Gertrude Weiss**, *Survey of Financial Characteristics of Consumers*, Washington 1966.

A. F. Shorrocks, "The Age-Wealth Relationship: A Cross-Section and Cohort Analysis," *Rev. Econ. Statist.*, May 1975, *57*, 155–63.

Carl S. Shoup, *Federal Estate and Gift Taxes*, Washington 1966.

James D. Smith, "White Wealth and Black People: The Distribution of Wealth in Washington, D.C. in 1967," in his *The Personal Distribution of Income and Wealth*, New York 1975.

J. Tobin, "Life-Cycle Saving and Balanced Growth," in William Fellner, ed., *Ten Economic Studies in the Tradition of Irving Fisher*, New York 1967.

M. E. Yaari, "Uncertain Lifetime, Life Insurance, and The Theory of the Consumer," *Rev. Econ. Stud.*, Apr. 1965, *32*, 137–50.

The Relationship between Relative Prices and the General Price Level: A Suggested Interpretation

By ALEX CUKIERMAN*

In a thought-provoking paper in this *Review*, Daniel Vining and Thomas Elwertowski examine empirically the relationship between the variability of the rate of inflation in the general level of prices, and the variance of the rate of change in relative prices. They find a positive association between these two variances and interpret this finding as a contradiction to a modern stochastic version of the neoclassical model as presented by Robert Lucas (1973). They state: "... thus, in contrast to many of his other conclusions, Lucas' remark in an otherwise extremely controversial paper, namely 'that there is no reason to expect τ to vary systematically with demand policies' (1976, p. 39), has gone utterly uncontested" (p. 706).[1] The implication is obviously that this type of model is inconsistent with the above mentioned finding. They go on to interpret their result in light of a recent paper by Robert Barro and interpret this paper as an "... effort to account for the observed dependence of heightened relative price change dispersion on general price change instability, relying upon a chain of causality running from general price level change instability to relative price change instability" (p. 707). They finally challenge empirical economists to discriminate between two hypotheses: "... i.e., to determine the direction of causality between individual price change dispersion and general price change instability" (p. 708).

I claim and demonstrate in this note that: 1) If correctly interpreted the type of many-markets stochastic model presented by Lucas is perfectly consistent with the finding that there is a positive association between individual price change dispersion and general price change dispersion. 2) It is wrong to interpret the Barro model as providing a rationale for "a chain of causality running from general price level change instability to relative price change instability" (Vining and Elwertowski, p. 707). It should rather be viewed as a conceptual framework in which *both* the variance of general price change and the variance of individual price change are influenced[2] by some common exogenous variances like the variance of aggregate excess demand shocks and the variance of relative excess demand shocks.[3] 3) Within a framework in which both the variance of general price change and the variance of relative price change are determined endogenously the question regarding the direction of causality between those two variances becomes ambiguous. If for example both variances increase because the variance of the (exogenous) rate of change in nominal income increases, it does not follow logically that either the variance of general price change causes the variance of relative price change, or vice versa. However, the question raised by Vining and Elwertowski does make sense if interpreted as a question concerning the direction of causality between *some* attributes of aggregate variability and *some* attributes of relative variability. Such an interpretation is suggested later.

I. Interpretation in Terms of Lucas' Model

If the Lucas model is interpreted as an economy in which price on each market is determined by a local market-clearing condition,[4] the following equilibrium relationship

*Carnegie-Mellon University, on leave from the University of Tel-Aviv.

[1] τ^2 is the variance of relative price changes in the Lucas model.

[2] Not always in the same direction.

[3] This is equally true for the correct interpretations of the Lucas model.

[4] Actually as pointed out by Lucas, only with such an interpretation does his 1973 paper make sense. For a

emerges between the variance of general price level σ^2, and the variance of the rate of change in nominal income σ_x^2,

$$\sigma^2 = \frac{\sigma_x^2}{(1+\gamma\theta)^2} \tag{1}$$

A similar relationship emerges between the variance of relative prices τ^2, and the variance of specific demand shocks σ_w^2,

$$\tau^2 = \frac{\sigma_w^2}{(1+\gamma\theta)^2} \tag{2}$$

where $\theta \equiv \sigma_w^2/(\sigma_x^2+\sigma_w^2)$ and γ is a positive constant. Inspection of σ^2 and τ^2 immediately reveals that they are not independent. They both depend on the exogenous variances σ_x^2 and σ_w^2. If either of these last two variances or both of them change over time, the variance of general price level change and the variance of relative price change will also change over time causing a definite systematic relationship to emerge between them. More formally, the (time dependent) variances of general price change and of individual price change will be given by[5]

$$\delta_{t+1}^2 \equiv Var\,[P_{t+1}-P_t] = E[P_{t+1}-P_{t+1}^* - (P_t - P_t^*)]^2 = \sigma_{t+1}^2 + \sigma_t^2 \tag{3}$$

$$\alpha_{t+1}^2 \equiv Var\,[P_{z,t+1}-P_{z,t}] = E(z_{t+1}-z_t)^2 = \tau_{t+1}^2 + \tau_t^2 \tag{4}$$

Here $P_{z,t}$, z_t, P_t, and P_t^* stand for the logarithm of price in market z, the percentage deviation of the price in market z from the general price level, the logarithm of the general price level, and the first moment of the logarithm of the general price level, all at time t.

Vining and Elwertowski find empirical evidence that suggests a positive association between α_t^2 and δ_t^2. As pointed out above, this positive association may be caused by changes in either σ_x^2 or σ_w^2 over time (or in both) which tend to change σ^2 and τ^2 in the same direction. The important question in this context is: Can such changes account for the observed positive association between α^2 and δ^2? Since, from (3) and (4), δ^2 depends only on a current and a lagged value of σ^2 and α^2 depends only on a current and lagged value of τ^2, this question is equivalent to the following question: Can independent exogenous changes in σ_x^2 and σ_w^2 cause a positive relationship between σ^2 and τ^2? The partial derivatives of σ^2 and τ^2 with respect to σ_x^2 and σ_w^2 are given in what follows.[6]

$$\frac{\partial\sigma^2}{\partial\sigma_x^2} = \frac{1+\gamma\theta+2\gamma\theta(1-\theta)}{(1+\gamma\theta)^3} > 0 \tag{5}$$

$$\frac{\partial\tau^2}{\partial\sigma_x^2} = \frac{2\gamma\theta^2}{(1+\gamma\theta)^3} > 0$$

$$\frac{\partial\sigma^2}{\partial\sigma_w^2} = -\frac{2\gamma(1-\theta)^2}{(1+\gamma\theta)^3} < 0 \tag{6}$$

$$\frac{\partial\tau^2}{\partial\sigma_w^2} = \frac{1+\gamma\theta(2\theta-1)}{(1+\gamma\theta)^3}$$

Expression (5) suggests that if only the variance of the rate of change in nominal income changes over time σ^2 and τ^2 will exhibit a positive association as found by Vining and Elwertowski. It is seen from (6) that a similar conclusion will follow for the case in which only σ_w^2 changes over time, if and only if $1+\gamma\theta(2\theta-1)<0$. Hence, if any changes that occur over time in σ_w^2 are dominated by changes in σ_x^2 over time, the positive association found between σ^2 and τ^2 is perfectly consistent with the Lucas model. The variance σ_x^2 may be changing relatively more than σ_w^2 over time, either because of frequent changes in policy or for other exogenous reasons beyond the control of policymakers.

II. Interpretation in Terms of Barro's Model

Except for a distinction between aggregate monetary shocks and real aggregate excess demand shocks and some other differences of structure, the Barro model and the Lucas

derivation of (1) and (2), see this author and Paul Wachtel, forthcoming.

[5]The extreme right-hand equalities in (3) and (4) follow from the assumption of no serial correlation in z_t and $P_t - P_t^*$. For further details, see the author and Wachtel, forthcoming.

[6]They are obtained by differentiating both (1) and (2) with respect to σ_x^2 and σ_w^2.

model are quite similar in spirit.[7] What is important for our purposes is that (as in the modified version of the Lucas model presented above) both the variance of relative price change and the variance of the general price level change are determined endogenously in the model. They are both determined by the values assumed by the variance of aggregate monetary shocks, the variance of aggregate real excess demand shocks, and the variance of relative excess demand shocks which are all considered exogenous.[8] Therefore, the direction of causality is not from general price change to relative price change, but rather from the above mentioned three exogenous variances to α_t^2 and δ_t^2. It is possible to investigate (as done for the Lucas model) under what conditions changes in the exogenous variances will cause positively related changes in α_t^2 and δ_t^2. It turns out that some restrictions have to be imposed on the parameters of the Barro model to yield a positive association between α_t^2 and δ_t^2. I do not present the full analysis for reasons of brevity since it is similar in substance (although not in details) to the previous analysis.

The main conclusion here is that Barro's model is no more a formulation of a model "... relying upon a chain of causality running from general price level change instability to relative price change instability" than it is a model relying upon a chain of causality running in the opposite direction. Rather a correct interpretation is that both types of price change instability are causally dependent on the variances of aggregate and relative excess demands shocks.

III. A Comparison to Parks' Model of Relative Price Variability

It may be instructive to digress for a little while to a recent alternative explanation of relative price variability which is due to Richard Parks. Parks presents empirical evidence that the variance of relative price change is related to some combination of the rate of change in real income and the size of unanticipated inflation. He interprets those results in terms of a multimarket framework that in some respects is more general and in others more restrictive than the one suggested here. On one hand he allows different demand and supply coefficients across different markets which is more general. On the other hand he restricts inflationary expectations and therefore the amount of unanticipated inflation to be identical across all markets.[9] In addition since his model is nonstochastic it does not yield itself easily to the investigation of the relationship between general and relative price level variability which is the main issue investigated by Vining and Elwertowski. A conceptual (stochastic) framework which would allow *both* different demand and supply coefficients as well as different expectations across markets is therefore highly desirable.

IV. Reformulation of the Causality Question

Both Lucas' and Barro's models assume that there is no relationship between the exogenous variances, and obtain as a conclusion the existence of *some* relationship between α_t^2 and δ_t^2. Since the Lucas variant accounts quite well for the observed positive relationship between α_t^2 and δ_t^2 (when their variation is caused mostly by variation in σ_x^2), it would seem that a more complicated theory of the kind hinted at by Vining and Elwertowski at the end of their paper is not necessary. However, there may be in reality feedbacks from a change in real relative variability to some component of aggregate variability. Such a feedback is probably implicit in Charles Schultze's theory of inflation,[10] in which a change in relative prices causes policy responses that lead to an increase in the general rate of inflation and possibly in its variance as well. The exact model to embody this notion has yet to be

[7]They are, however, used to investigate different issues.

[8]In Barro's notation those three variances are σ_m^2, σ_v^2, and σ_ϵ^2, respectively; see his pp. 5–7.

[9]For an investigation of the relationship between the variance of relative price change and the variance of inflationary expectations within a framework that allows nonuniform expectations across markets, see the author and Wachtel (1978).

[10]See also Robert Solow, pp. 62–65

worked out. However, it should be clear that within the present framework such an hypothesis should be represented as a chain of causality running from σ_w^2 (or from τ^2) to σ_x^2 rather than to σ^2.[11] Such a theory will probably also yield a positive relationship between σ^2 and τ^2. I would therefore rephrase the challenge raised by Vining and Elwertowski at the end of their paper in the following way: Is the positive relationship between α_t^2 and δ_t^2 explainable only in terms of a variation in σ_x^2, given that σ_w^2 and σ_x^2 are not related, or does it also point to the existence of some relationship between σ_w^2 and σ_x^2 not accounted for in either the Lucas or the Barro model?

[11]Obviously once σ_x^2 changes, it will affect both σ^2 and τ^2 causing further changes in σ_x^2

REFERENCES

R. Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

A. Cukierman and P. Wachtel, "Differential Inflationary Expectations and the Variability of the Rate of Inflation: Some Theory and Evidence," *Amer. Econ. Rev.*, forthcoming.

——— **and** ———, "Relative Price Variability and Non-Uniform Inflationary Expectations," no. 78-13, Center Applied Econ., New York Univ., Aug. 1978.

R. E. Lucas, Jr., "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

———, "Econometric Policy Evaluation: A Critique," *J. Monet. Econ.*, 1976, suppl., *2*, 19–46.

R. Parks, "Inflation and Relative Price Variability," *J. Polit. Econ.*, Feb. 1978, *86*, 79–96.

C. Schultze, "Recent Inflation in the U.S.," study paper no. 1 for *Joint Economic Committee*, Washington 1959.

R. Solow, "The Intelligent Citizen's Guide to Inflation," *Publ. Interest*, Winter, *38*, 30–66.

D. R. Vining and T. C. Elwertowski, "The Relationship Between Relative Prices and the General Price Level," *Amer. Econ. Rev.*, Sept. 1976, *66*, 699–708.

Exchange Market Pressure in Postwar Brazil: An Application of the Girton-Roper Monetary Model

By MICHAEL CONNOLLY AND JOSÉ DANTAS DA SILVEIRA*

This study applies Lance Girton and Don Roper's (hereafter G-R) monetary model of exchange market pressure to the postwar Brazilian monetary experience. The model was designed specifically for the Canadian managed float during the period 1952–62. The object of their model is to explain what they term "exchange market pressure"; that is, the pressure on foreign exchange reserves *and* the exchange rate when there exists an excess of domestic money supply over money demand in a managed floating exchange rate regime. The basic theoretical proposition is that any such excess supply of money can be relieved by an exchange depreciation, a loss in foreign reserves, or, in the context of a managed float, by some combination of the two. In this sense, the G-R managed float model used here is firmly rooted in the modern monetary approach to exchange rates and the balance of payments.[1]

Brazil provides a particularly good example for testing this approach, not only because it is in many senses a unique example of a postwar managed float system, but also because it can be treated as a "small, open" economy in the sense that world prices and monetary conditions faced by Brazil are taken as given. This particularly suits the purpose of most modern monetary models which make this assumption and obviates the problems of monetary dependence and neutralization dealt with in the pioneering G-R paper. Specifically, the small-country assumption permits us to devise a simple one-country equation of managed floating which depends upon four essential ingredients: 1) money demand, 2) money supply, 3) purchasing power parity, and 4) monetary equilibrium.[2] Furthermore, in Brazil a much greater proportion of exchange market pressure was absorbed by exchange rate depreciation than in the Canadian case where changes in reserves were large relative to exchange rate movements. In short, postwar Brazil provides a singularly good opportunity to test the monetary model of exchange market pressure.

Section I briefly states the essential elements of the monetary model, and derives the equation to be tested for the Brazilian experience from 1955 to 1975. Section II reports empirical results for the exchange market pressure model, and Section III examines the applicability of the relative version of purchasing power parity for the time period considered. Section IV summarizes the results and discusses the merits of the monetary approach in light of the Brazilian experience.

I. A Simple Version of the G-R Exchange Pressure Model

As noted in Johnson, "A proper test of the monetary approach must be essentially a test of the stability of the demand for money (in Friedman's terminology)," (1977, p. 263) and, accordingly, we postulate a stable demand for money which depends upon domestic prices P, and the level of real permanent income Y:

$$L = kPY \tag{1}$$

where k is the fraction of yearly income that firms and households wish to hold in the form of money balances. For simplicity, this fraction is assumed constant.

*University of South Carolina and the Federal University of Ceará, Brazil, respectively. We are indebted to Don Roper for helpful comments.

[1]See Rudiger Dornbusch, Jacob Frenkel and Harry Johnson, Johnson (1972), and Robert Mundell.

[2]The complete monetary model of exchange market pressure is found in G-R.

Second, we specify the money supply process. Following Mundell, we may consolidate private and central banks into a single banking system whose liabilities, the money stock M, are matched by the sum of its foreign and domestic assets, denoted by R and D, respectively, in the following identity:

$$M \equiv R + D \tag{2}$$

This identity states that changes in the money stock are either from foreign or domestic sources, for example, a change in foreign reserves via the balance of payments or a change in domestic credit extended by the consolidated banking system.

Third, in accordance with the small, open country assumption, we suppose that domestic prices P reflect foreign prices P^* via the exchange rate E, or

$$P = EP^* \tag{3}$$

Finally, it simplifies matters greatly to assume that the money stock in existence adjusts rapidly to the quantity demanded, either by a deficit running down the money stock or by an exchange depreciation, or by some combination of the two, so that monetary equilibrium holds:

$$M = L \tag{4}$$

Substituting (3) into (1), and (1) and (2) into (4), differentiating logarithmetically, and expressing the result in terms of percentage changes, we have after manipulation an equation corresponding to G-R's exchange market pressure:

$$r + e = -d + p^* + y \tag{5}$$

where r is the change in foreign reserves (or the balance of payments) as a proportion of the money supply,[3] e is the percentage appreciation (if positive) of the Brazilian cruzeiro, d is the change in domestic credit as a proportion of the money stock, p^* is the world rate of inflation, and y the rate of growth of permanent income. This equation states that *an increase in the rate of growth of domestic credit, for a given rate of growth of world prices and permanent income, will result in an equiproportionate loss in reserves with no change in the exchange rate, or an equiproportionate depreciation of the cruzeiro, or some combination of the two.*

II. Empirical Results

In the immediate postwar period from 1947 to 1953, rigid exchange controls were in effect in Brazil. This was followed by an "exchange auction" system from October 1953 to March 1961, which was characterized by an official exchange rate depending upon the category of transaction involved, an exchange tax, and an auction certificate premium, which amounted to a system of multiple exchange rates. In March 1961 a single rate was introduced and controls were further relaxed until in August 1968, a "mini-devaluation" system was introduced and has stayed in effect until the present. For purposes of empirical testing, we eliminated from consideration the early rigid control period and estimated equation (5) for a twenty-one year period, 1955–75; and then a shorter subperiod of fourteen years, 1962–75. These periods were characterized by relatively unfettered exchange market conditions that more closely correspond to the theoretical model. Insofar as possible, the exchange rate used in estimation reflected actual market rates in effect rather than simply official rates.[4] The definition of domestic credit corresponds to money narrowly defined, less net foreign reserves of the consolidated banking system. The change in the *U.S.* Wholesale Price Index was used for p^*, and changes in a three-year moving average of real gross domestic product used for y. The results are reported in equations 1 and 2 of Table 1.[5]

[3]The Banco Central do Brasil (*BCB*) reports changes in the net foreign reserves of the consolidated banking system for its balance of payments, which permits us to deflate changes in reserves by the money stock. Girton and Roper use Canadian Official International Reserves and consequently deflate reserve changes by base money. Also, the Brazilian cruzeiro exhibited much larger movements than the Canadian dollar, so deflating by the money stock rather than by base money has the advantage of making measures of exchange rate pressure absorbed by r vs e comparable in magnitude (see the Appendix).

[4]See the Appendix for discussion of sources.

[5]For a full treatment of the Brazilian exchange rate system, see Don Huddle, Alexander Kafka, and Eugenio Gudin.

TABLE 1

Regression Equation	Period	Dependent Variable	Estimated Coefficients of the Independent Variables d	p^*	y	R^2 / F	$D.W.$ / Rho	SSR / SER
1	1955–75	$r+e$	−1.009	1.285	1.268	0.68	2.22	0.81
			(−7.417)	(1.266)	(1.262)	17.68	−0.11	0.22
2	1962–75	$r+e$	−1.014	1.207	1.456	0.91	2.00	0.13
			(−13.092)	(2.042)	(2.446)	50.15	−0.12	0.11
3	1955–75	r	−0.428	−0.131	2.277	0.55	1.56	0.22
			(−3.752)	(−0.169)	(2.410)	10.52	0.62	0.11
4	1962–75	r	−0.178	−0.675	2.308	0.48	1.56	0.14
			(−1.469)	(−0.848)	(2.709)	4.65	0.30	0.12

Note: As in G-R, the Cochrane-Orcutt iterative technique was used to adjust for serial correlation with *Rho* being the estimated auto-regression coefficients The numbers in parentheses below the estimated coefficients indicate t-values; R^2 = coefficient of determination; F = F-statistic, $D\ W$ = Durbin-Watson statistic; SSR = sum of squared residuals, SER = standard error of the regression. Other symbols are defined in the text.

While the evidence on domestic credit in both periods is consistent with the monetary model of exchange rate pressure, the price and income coefficients are not significant from 1955 to 1975, but are from 1962 to 1975. That is, the predictive performance is, as expected, stronger in the later period. In all cases, however, the estimated coefficients correspond to their hypothesized theoretical values of minus unity for d, and plus unity for p^* and y. Indeed, a joint F-test restricting *all* coefficients to equal the value hypothesized by the monetary model in equation (5) compared to the unrestricted regressions reported in Table 1 gives an F-statistic of .25 for 1955–75, and .73 for 1962–75. In both cases, the simple monetary model of exchange rate pressure would not be rejected at the 1 percent level.

As a further test of the exchange market pressure model, we duplicated Nicholas Sargen's tests. He estimated regressions for Canada from 1952 to 1974, and Australia, Germany, Japan, and the United Kingdom from 1962 to 1975, and found that with the exception of Canada, using r as the sole dependent variable rather than the composite variable $r+e$ gave a better overall fit. We performed the same test and report the results in equations 3 and 4 of Table 1. They are somewhat poorer for the entire period and dramatically worse for the later, when there were fewer exchange restrictions.

As a final test of the exchange market pressure model, we tested the hypothesis that the implied tradeoff between reserve losses and exchange depreciation for the monetary authorities was one-to-one; for example, that the measure $r+e$ is not sensitive to its composition between exchange rate and reserve changes. To do this, we included the variable $Q = (e-1)/(r-1)$ on the right-hand side of the exchange market pressure equation (5) in regressions from 1955–75 and 1962–75. The variable Q is a good measure of the way in which the monetary authorities absorb exchange market pressure because the more they let pressure be alleviated by depreciation relative to reserve losses, the greater is Q.[6] The simple ratio e/r used by G-R does not have this desirable monotonic property since it is discontinuous for values of r equal to zero, which is empirically important for Brazil since surpluses turned into deficits and deficits into surpluses frequently over the period (see the appendices for values of r and e, and more discussion on this point).

[6]A sufficient condition for this is that the balance of payments not have a surplus equal to its total money stock, and that the currency not double in price in any one year. That is, both r and e fall short of unity.

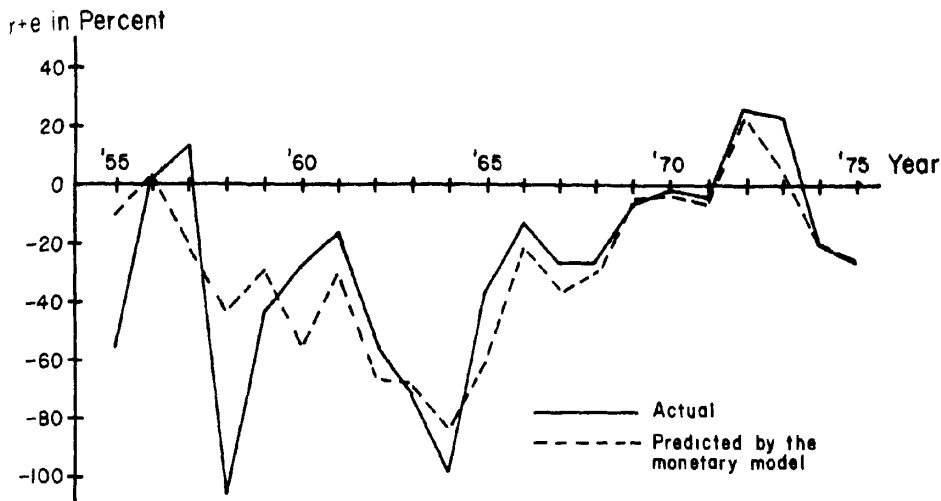



FIGURE 1. BRAZILIAN EXCHANGE MARKET PRESSURE, 1955–75

The results were, from 1955 to 1975 (t-values in parentheses)

$$(6)\quad r + e = \underset{(-3.40)}{-.85d} + \underset{(1.13)}{1.20p^*} + \underset{(1.63)}{2.78y} - \underset{(-1.09)}{.13Q}$$

$R^2 = .70$; $\rho = -.04$; $S.E. = .22$; $D.W. = 2.02$

and from 1962 to 1975 (again with t values in parentheses)

$$(7)\quad r + e = \underset{(-6.17)}{-1.12d} + \underset{(2.17)}{1.40p^*} + \underset{(.43)}{.59y} + \underset{(.69)}{.07Q}$$

$R^2 = .91$; $\rho = -.20$; $S.E. = .12$; $D.W. = 2.09$

The variable Q is not significant, and other coefficients remain essentially the same (apart from the now low and insignificant income elasticity of demand for money in equation (7)), which indicates that the measure of exchange market pressure is not sensitive to its composition. This is consistent with G-R's finding and their suggestion that the variable $r + e$ can be used to determine the level of exchange intervention needed to achieve exchange rate targets.[7]

In short, taking into account Sargen's test along with simple G-R tests, we find that the monetary approach explains exchange market pressure in postwar Brazil to a rather remarkable extent. This is reflected in Figure 1 which plots as a solid line the actual movements in the composite variable $r + e$, alongside the dashed line for movements predicted by the theoretical model $(-d + p^* + y)$. This graph differs only slightly from that of G-R for Canada (p. 544) in that they compare the fitted or estimated prediction with the actual outcome, rather than a theoretical prediction. As is clear from inspection of Figure 1, the monetary model predicts better in the later years than in earlier ones. Reasons for this are taken up in the next section.

III. Purchasing Power Parity

As a result of the more stringent exchange control system in the early postwar years,

[7]See G-R, p. 545. It might be noted that Johnson (1972, p. 1572) derived an identical formula for the movement in the exchange rate as a function of both domestic credit and exchange intervention policies for a

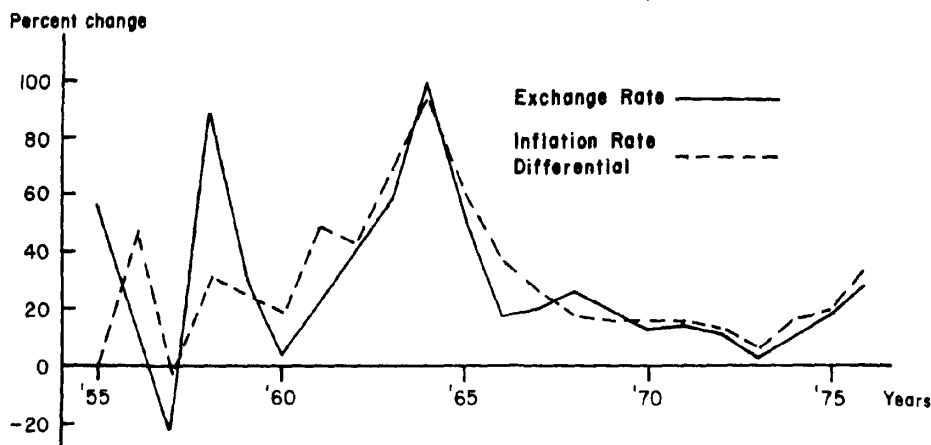



FIGURE 2 PURCHASING POWER PARITY IN BRAZIL, 1955–76

1955–61, the Brazilian economy was somewhat sheltered from foreign economic forces and was able to pursue a more independent domestic credit policy than typically the case for a small open economy. The consequence of this insulation was that the domestic rate of inflation in Brazil deviated substantially from the rate of depreciation of the cruzeiro plus the world inflation rate, the relationship predicted by purchasing power parity in equation (3).[8] However, from 1962 to 1975, purchasing power parity holds very well, and, as a consequence, the monetary model performs better in a situation in which the Brazilian economy was more "open" in the strict sense of the term.

Figure 2 plots the percentage change in the cruzeiro price of the dollar as a solid line and the percentage change in the Brazilian consumer price index less the percentage change in the *U.S.* Consumer Price Index as a dashed line from 1955 to 1976. Clearly, the association between the two is much closer from 1962 to 1976 than from 1955 to 1962. In light of the exchange rate restriction in the early years, it seems that purchasing power parity stands up rather well. This is quite important since it is an integral part of the transmission mechanism of the monetary approach which is, at the moment, the subject of considerable controversy.[9]

IV. Summary Remarks

Postwar Brazil provides an excellent example of an exchange market in which for over two decades exchange market pressure has been alleviated by a combination of reserve changes and exchange depreciation. The simple monetary model of exchange rate pressure tested here performs fairly well in the 1955–75 period and very well during 1962–75, in explaining movements of reserves and the exchange rate. Its poorer performance in the early period up to 1961 resulted in part from greater exchange restrictions that were

given world rate of inflation and growth in the home economy.

[8]Note that we need not take up here the question of direction of causation, that is, *from* price movements to exchange movements or the reverse. This is in any case an incorrect way of viewing a general equilibrium problem where both are determined simultaneously. See Frenkel on this.

[9]On the side favorable to the so-called "law of one price" are Hans Genberg, Donald McClosky and Richard Zecher, while Robert Dunn, Michael Bordo and Ehsan Choudri come out against, and J. David Richardson reports mixed results. In the context of devaluation, Michael Connolly and Dean Taylor find that it predicts poorly for the year or so before and after devaluation during the adjustment process, but fairly well over two-year periods.

eased in March of that year. The purchasing power parity relationship also holds up well, particularly after 1961.

APPENDIX: DATA AND DATA SOURCES

Net flows of foreign assets as a proportion of the money supply (r), were obtained by multiplying the balance of payments in *U.S.* dollars by the average exchange rate, and dividing by the money supply of the previous year. Data for the balance of payments, up to 1973, were drawn from bulletins of the *BCB* issues of December 1972, April and June 1973, and December 1974. Brazil reports the net change in foreign reserves of both monetary authorities and commercial banks as the balance of payments, so that the balance-of-payments figures strictly correspond to changes in net foreign reserves of the consolidated banking system, as in equation (2). Data for 1974 and 1975 were drawn from the 1975 Annual Report of the *BCB*. Data for the money supply (currency held by the public plus demand deposits) were drawn from *Conjuntura Econômica* (*CE*) of April 1977.

The average exchange rate used for each year represents a simple average of monthly averages. Up to August 1968 these monthly averages were calculated by the Getúlio Vargas Foundation, an official source for Brazilian data, and represent rates "applicable to imports of a general character effected by private persons" (*CE*, Sept. 1963, p. 75). See *CE* of September 1963 and October 1968. After August 1968 the corresponding averages were calculated by us from selling rates reported in bulletins of the *BCB*.

The rates of expansion of domestic credit as a proportion of the money supply d were obtained simply by subtracting r from the rate of expansion of the money supply.

The percent change in a three-year moving average of Gross Domestic Product (*GDP*) at 1970 prices was used as a proxy of the rate of growth of Brazilian permanent income. For the last observation (1975) on permanent income, however, we used the actual value of *GDP*. Data on *GDP* were drawn from the Supplement of *CE* of November 1972 (from 1953 to 1962) and from the International Financial Statistics (*IFS*) of May 1977 for the remaining period.

Finally, data on price indexes were drawn from the *IFS* of May 1976 and May 1977.

To see clearly that e/r is not an appropriate variable for the purpose of measuring the relative burden of adjustment to exchange market pressure, consider the following observations:

	e	r	$r+e$	e/r	Q
1	−0.6	0.3	−0.3	−2.0	2.28
2	−0.4	0.1	−0.3	−4.0	1.56
3	−0.2	−0.1	−0.3	2.0	1.09
4	−0.1	−0.2	−0.3	0.5	0.92

Notice that while "exchange market pressure" is less absorbed by devaluation—going from 60 percent in observation 1 to 10 percent in observation 4—the ratio e/r declines from the first to the second observation but, due to its discontinuity for r equal to zero, increases from the second to the third observation. The variable Q however declines steadily when the same amount of exchange market pressure is absorbed relatively less by exchange rate changes than reserve changes.

TABLE A—BRAZILIAN EXCHANGE MARKET PRESSURE 1955–75

Year	e	r	$e+r$
1955	−.58	.01	−.57
1956	−.14	.14	.00
1957	.22	−.09	.14
1958	−.90	−.19	−1.09
1959	−.33	−.11	−.45
1960	−.04	−.24	−.28
1961	−.21	.06	−.16
1962	−.39	−.17	−.56
1963	−.59	−.12	−.72
1964	−1.00	.00	−1.00
1965	−.53	.18	−.36
1966	−.17	.05	−.12
1967	−.20	−.07	−.27
1968	−.27	.01	−.27
1969	−.20	.13	−.07
1970	−.13	.11	−.02
1971	−.15	.10	−.05
1972	−.12	.38	.26
1973	−.03	.27	.24
1974	−.11	−.09	−.20
1975	−.20	−.07	−.27

Note: Minor inconsistencies due to rounding.

REFERENCES

M. Bordo and E. Choudri, "The Behavior of the Prices of Traded and Non-Traded Goods: The Canadian Case 1962–74," *Carleton Econ. Pap.*, Carleton Univ., Dec. 1976.

M. Connolly and D. Taylor, "Testing the Monetary Approach to Devaluation in Developing Countries," *J. Polit. Econ.*, Aug. 1976, *84*, 849–59.

R. Dornbusch, "Devaluation, Money and Non-Traded Goods," *Amer. Econ. Rev.*, Dec. 1973, *63*, 871–80.

R. Dunn, "Flexible Exchange Rates and Oligopoly Pricing: A Study of Canadian Markets," *J. Polit. Econ.*, Jan. 1970, *78*, 140–51.

J. Frenkel, "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scand. J. Econ.*, May 1976, *78*, 200–25.

——— **and H. Johnson,** "The Monetary Approach to the Balance of Payments: Essential Concepts and Historical Origins," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976, 21–45.

H. Genberg, "Aspects of the Monetary Approach to Balance of Payments Theory: An Empirical Study of Sweden," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976, 298–326.

L. Girton and D. Roper, "A Monetary Model of Exchange Market Pressure Applied to the Postwar Canadian Experience," *Amer. Econ. Rev.*, Sept. 1977, *67*, 537–48.

E. Gudin, "Multiple Exchange Rates—The Brazilian Experience," *Econ. Int.*, Aug. 1956, *9*, 501–9.

D. Huddle, "Balanço de Pagamentos e Contrôle de Câmbio no Brasil, Diretrizes Políticas e História: 1946–1954," *Rev. Brasileira Econ.*, Jan. 1964, *18*, 5–40.

H. Johnson "The Monetary Approach to Balance of Payments Theory," *J. Finan. Quant. Anal.*, Mar. 1972, *7*, 1555–72.

———, "The Monetary Approach to the Balance of Payments: A Non-Technical Guide," *J. Int. Econ.*, Aug. 1977, *7*, 251–68.

A. Kafka, "The Brazilian Exchange Auction System," *Rev. Econ. Statist.*, Aug. 1956, *38*, 308–22.

D. McClosky and R. Zecher, "How the Gold Standard Worked, 1880–1913," in Jacob A. Frenkel and Harry G. Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976, 357–86.

Robert Mundell, *Monetary Theory*, Pacific Palisades 1971, 86–97.

J. Richardson, "Some Empirical Evidence on Commodity Arbitrage and the Law of One Price," unpublished paper, Univ. Wisconsin-Madison, Mar. 1977.

N. Sargen, "An Empirical Analysis of a Monetary Model of Exchange Rate and Reserve Fluctuations," unpublished paper, Federal Reserve Bank San Francisco, Nov. 1975.

Banco Central do Brasil (*BCB*), *Boletim do Banco Central do Brasil*, Rio de Janeiro, various issues.

———, *Relatório Anual*, Rio de Janeiro, 1975.

Fundação Getúlio Vargas, *Conjuntura Econômica* (*CE*), Rio de Janeiro, various issues and supplements.

International Monetary Fund, *International Financial Statistics* (*IFS*), Washington, various issues.

Income and Substitution Effects in the Two-Sector Open Economy

By GIORA HANOCH AND MORDECHAI FRAENKEL*

The analysis of a small, open economy with two sectors, tradable and nontradable goods, has important implications for small countries operating under a fixed exchange rate, or even a managed floating rate. In particular, the question arises, what are the conditions for a successful devaluation, intended to reduce a current deficit in the balance of payments?

In a recent valuable contribution to this topic, Michael Bruno analyzes the effects of changes in the real exchange rate (the price of tradables relative to nontradables) q on net supplies of the two goods in the private sector. He reaches the conclusion that "the net supply of nontradables unambiguously decreases with an increase in q" (p. 569), i.e., $S_{0q} < 0$.[1] On the other hand, he states that "For tradables, the substitution and income effects work in the opposite directions, and the result is therefore more ambiguous" (p. 569). He also gives a necessary and sufficient condition (equation (7)) for the effect of q on the net supply of tradables to have the expected sign $S_{1q} > 0$.

In this note we show that the substitution effects in the production as well as the consumption of both types of goods do have the expected signs; however, a sizable income effect may produce irregular results, so that the total effect of a devaluation on the net supply of either good may have the "wrong" sign, and this depends crucially on the gap between the relative shares of tradables in output and in consumption.

It is also shown that Bruno's unambiguous effect of q on the net supply of nontradables is the result of an unnecessarily restrictive assumption, namely that the consumption demand for tradables is price elastic ($\eta_1 > 1$).

Analyzing the general case (assuming, however, that both goods are normal in consumption), we derive the conditions under which both supplies display the expected responses simultaneously ($S_{1q} > 0$ and $S_{0q} < 0$); that is, when the combined substitution effect in production and consumption is dominant. Finally, we point out the policy implications of the other two possible cases (when S_{1q} and S_{0q} have the same sign), which Bruno and others ignore, and which may arise if the shares of tradables in consumption and in production differ too widely in one direction or the other. In one case ($S_{1q} < 0$ and $S_{0q} < 0$), the analysis indicates a "reverse policy"—i.e., a *revaluation*, accompanied by a reduction in the absorption rate s, or an increase in government spending on nontradables G_0.

In the other case ($S_{1q} > 0$ and $S_{0q} > 0$), the short-run market equilibrium is unstable; however, the government may restore stability (and the expected responses) by manipulating its three policy tools: the nominal exchange rate e, public expenditure on nontradables G_0, and the absorption rate s.

Using Bruno's assumptions[2] and notation we have the following definitions ($i = 0,1$):

$q = p_1/p_0 = e/p_0$ is the real exchange rate, where p_0 is the unit price of nontradable goods, and p_1 is the domestic price of tradables, equal to the nominal exchange rate e (since the international unit price p_1 equals 1, by normalization).

*The Hebrew University and the Bank of Israel, respectively. We are indebted to Michael Bruno and Elhanan Helpman for very helpful comments.

[1] $S_{0q} >$ is a misprint in Bruno, p 569.

[2] The present formulation follows Bruno in excluding an explicit treatment of money and monetary policy. See, for example, Rudiger Dornbusch for such an explicit treatment (in a world with two countries and no active government). However, the explicit introduction of money would not change the main aspects of the model (as explained in Bruno, pp. 569; 571, fn. 13), and our main conclusions.

$C_i(C,q)$ are demand functions, satisfying

$$\frac{\partial C_i}{\partial C} = C_{ic} \geq 0 \text{ (normality).}$$

$X_i(q)$ are supply functions, satisfying in equilibrium

$$(1) \quad X_{0q} = -qX_{1q} < 0$$

(by convexity of the production set).

$S_i(q) = X_i(q) - C_i(C,q)$ are net supplies of the private sector, where C is total consumption, satisfying by definition

$$C = C_0 + qC_1 = (1-s)Y = (1-s)(X_0 + qX_1)$$

where Y is total output. (The "saving" or absorption rate s is an exogenous policy parameter, assumed to be constant throughout most of this discussion.)

The demand derivatives satisfy the following equations:

$$(2) \quad C_{0c} + qC_{1c} = 1$$

$$(3) \quad C_{0q} + qC_{1q} = -C_1$$

$$(4) \quad C_{iq} = C^*_{iq} - C_1 C_{ic} \quad (i = 0, 1)$$

Equation (4) is Slutsky's equation, where C^*_{iq} are compensated price effects, or pure substitution effects, satisfying $C^*_{0q} > 0$, $C^*_{1q} < 0$ and

$$(5) \quad C^*_{0q}/C^*_{1q} = -q$$

(by optimality of demands).

First we note that the unambiguous sign $S_{0q} < 0$ in Bruno is a result of his restrictive assumption $C_{0q} \geq 0$ (p. 568), which is equivalent to

$$\eta_1 = -\frac{qC_{1q}}{C_1} \geq 1, \text{(by equation (3));}$$

that is, the consumption demand for tradables is price elastic. Since this is an empirical statement, there is no a priori reason for assuming it.

Secondly, we reformulate Bruno's condition for $S_{1q} > 0$; that is, for an increase in the real exchange rate to increase the net supply of tradables. Differentiating $S_1(q)$ we get

$$(6) \quad S_{1q} = X_{1q} - C_{1q} - C_{1c}\frac{\partial C}{\partial q} =$$

$$X_{1q} - C_{1q} - C_{1c}\frac{\partial}{\partial q}\{(1-s)(X_0 + qX_1)\} = X_{1q} - C_{1q} - C_{1c}(1-s)X_1$$

using equation (1). (Equation (6) is equivalent to Bruno's equation (7), which is given in terms of elasticities.)

Substituting from equation (4) for C_{1q} separates the substitution and income effects:

$$S_{1q} = X_{1q} - C^*_{1q} - C_{1c}\{(1-s)X_1 - C_1\}$$

where the last term is the total income effect on the net supply of tradables, which combines the conventional income effect of the change in q on demand for tradables $(-C_1 \cdot C_{1c})$ with the effect of the increase in income through production $(1-s)X_1 \cdot C_{1c}$.

Hence, $S_{1q} > 0$ if and only if the combined positive substitution effect $X_{1q} - C^*_{1q}$ is dominant; i.e., if

$$(7) \quad \frac{X_{1q} - C^*_{1q}}{C_{1c}} > (1-s)X_1 - C_1 =$$

$$S_1 - sX_1 = \frac{C}{q}\left\{\frac{qX_1}{Y} - \frac{qC_1}{C}\right\} = \Delta$$

Equation (7) shows that if $\Delta < 0$, then S_{1q} is positive, since $(X_{1q} - C^*_{1q})/C_{1c}$ is positive. However, if $\Delta > 0$, then $S_{1q} > 0$ if and only if the marginal propensity to consume tradables (C_{1c}) is sufficiently small relative to the combined substitution effect $(X_{1q} - C^*_{1q})$. The income effect on tradables indeed "works in the opposite direction" (Δ is positive) if and only if consumption is less intensive in tradables than is production.

Next, a similar analysis gives the condition for $S_{0q} < 0$, that is, for a real devaluation to decrease the net supply of local nontradable goods. Differentiating $S_0(q)$ we get

$$S_{0q} = X_{0q} - C_{0c}(1-s)X_1 - C_{0q}$$

Substituting the Slutsky equation (4) for C_{0q} gives

$$S_{0q} = X_{0q} - C^*_{0q} - C_{0c}\{(1-s)X_1 - C_1\} = X_{0q} - C^*_{0q} - C_{0c} \cdot \Delta$$

where Δ is defined in equation (7). The condition for $S_{0q} < 0$ is therefore

$$(8)\qquad \frac{X_{0q} - C^*_{0q}}{C_{0c}} < \Delta$$

But $X_{0q} - C^*_{0q} = -q(X_{1q} - C^*_{1q}) < 0$ (using equations (1) and (5)); hence, if Δ is positive, then $S_{0q} < 0$. But if $\Delta < 0$ (i.e., consumption is more intensive in tradables than is output) then by equation (8), $S_{0q} < 0$ if and only if the marginal propensity to consume nontradables (C_{0c}) is sufficiently small relative to the total substitution effect $-(X_{0q} - C^*_{0q}) = q(X_{1q} - C^*_{1q})$.

Combining equations (7) and (8) gives a necessary and sufficient condition for both effects to have the expected signs, namely $S_{1q} > 0$ and $S_{0q} < 0$:

$$(9)\qquad \frac{X_{0q} - C^*_{0q}}{C_{0c}} < \Delta < \frac{X_{1q} - C^*_{1q}}{C_{1c}}$$

Thus, both effects have these signs simultaneously if and only if Δ lies within the interval specified in equation (9), which contains $\Delta = 0$. The end points of this interval are determined by the ratio of the combined substitution effect to the income effect in consumption of tradables and nontradables, respectively. As long as the discrepancy between the shares of tradables in consumption and in production is not too wide, Δ is small in absolute size, condition (9) is likely to hold, and both $S_{1q} > 0$ and $S_{0q} < 0$.

The simple case $\Delta = 0$, i.e., when $qC_1/C = qX_1/Y$ (or $C_0/C = X_0/Y$), implies that under equal shares both effects have the expected sign (assuming normality), regardless of the size of the relevant demand and supply derivatives.

We now analyze the two cases where equation (9) does not hold. First, if

$$\Delta > \frac{X_{1q} - C^*_{1q}}{C_{1c}} > 0$$

then $S_{0q} < 0$ but also $S_{1q} < 0$. That is, the market for tradables is somewhat irregular, since a devaluation that increases the real exchange rate q leads to a decrease in excess supply S_1, and increases the deficit on current account (or decreases the surplus). The indicated exchange rate policy for reducing a deficit is therefore to *revalue* (reduce q or the nominal rate e).

However, since $S_{0q} < 0$, such a revaluation would increase the excess supply of nontradables, and deflationary pressures would be operating to reduce p_0. Since $q = e/p_0$, the real exchange rate q and the deficit would return in the long run to their former levels, unless government accompanies the revaluation with a decreased absorption through a *decrease* in s.[3] Thus the indicated policy changes in this case are the opposite of those analyzed in Bruno (namely $S_{0q} < 0$ and $S_{1q} > 0$), with regard to both q and s.

Now consider the second case

$$\Delta < \frac{X_{0q} - C^*_{0q}}{C_{0c}} < 0$$

implying $S_{1q} > 0$ but also $S_{0q} > 0$. Expressing Δ in terms of nontradables gives

$$(10)\qquad \Delta = (1 - s)\cdot\frac{Y - X_0}{q} - \frac{C - C_0}{q}$$
$$= \frac{1}{q}\{C_0 - (1 - s)X_0\}$$
$$= \frac{1}{q}(sX_0 - S_0)$$

If government purchases of nontradables are given at G_0,[4] then the market clears by adjusting the price p_0, so that *ex post* the equality $S_0 = G_0$ holds, and government purchases equal the private excess supply of nontradables S_0. But if $S_{0q} > 0$, the short-run equilibrium in this market is unstable, since an increase in q (given G_0 and s) is equivalent to a decrease in the own price of nontradables p_0, which is accompanied by a further increase in excess supply S_0.

Equation (8) may therefore be viewed as a restriction on the three policy parameters q, s, and G_0, that are required for short-run stability of the market for nontradables:

$$(11)\qquad \frac{X_{0q} - C^*_{0q}}{C_{0c}} < \frac{1}{q}\{sX_0(q) - G_0\} = \Delta$$

Condition (11) can only be violated, if $G_0 > sX_0$; that is, when the real exchange

[3]Alternatively, the government may increase its purchases of nontradables G_0.

[4]The size of G_0 is bounded on both sides: (a) G_0 is nonnegative if government does not hold stocks of nontradables; (b) G_0 cannot exceed total production X_0.

rate q is "too small" given s and G_0; or alternatively, when government demand for nontradables G_0 is "too large" give s and q.

Note that the point $\Delta = 0$ can always be reached if the production possibility curve is well behaved (i.e., $-\partial X_0/\partial X_1$ varies continuously from zero to infinity). Therefore, if $\Delta < 0$ the government can always ensure stability (i.e., $S_{0q} < 0$), by manipulating government purchases of nontradables G_0 and the absorption rate s, so as to decrease the absolute size of Δ and satisfy equation (11).

Note also that the case $S_{1q} < 0$ does not imply an analogous instability in the market for tradables, since the world supply of tradables is perfectly elastic, and government net purchases G_1 may exceed—even in the short run—the local excess supply S_1, with the difference $G_1 - S_1$ reflected in the deficit on current account. Finally, the existence of a deficit ($G_1 - S_1 > 0$) does not necessarily imply a higher intensity of tradables in consumption than in production (i.e., $\Delta = S_1 - sX_1 < 0$), since $G_1 \neq sX_1$.

REFERENCES

M. Bruno, "The Two-Sector Open Economy and the Real Exchange Rate," *Amer Econ. Rev.*, Sept. 1976, *66*, 566–77.

R. Dornbusch, "Devaluation, Money, and Nontraded Goods," *Amer. Econ Rev.*, Dec. 1973, *63*, 871–80.

Should Government Subsidize Risky Private Projects?: Comment

By MARION B. STEWART*

In a recent article in this *Review*, Joram Mayshar has argued that a central government should indeed subsidize a privately owned firm's investments in risky projects, if (a) capital markets are imperfect, and (b) the firm is subject to an income tax.

The intuitive justification for Mayshar's conclusion is as follows. In an economy with imperfect capital markets, privately owned firms will be unable to spread the risk associated with speculative investments over a sufficiently large number of investors to drive the costs of risk bearing to zero, and hence will rationally choose an investment strategy which differs from that which would be chosen by a risk-indifferent firm. This private investment strategy will also differ from that which would be chosen by a central government, since the government can render the risk-bearing costs of an independent project negligible by spreading the risk over the entire population of taxpayers; but this is not in itself a justification for any government action. The risk-bearing costs of the private project are real costs to the economy, and there is no presumption that the firm which stands to benefit from the investment should not be the entity which bears those costs—in this case by choosing a risk-reducing investment schedule which will on the average also reduce profits.

The justification for government action comes from an unexpected source: the corporate income tax. As Evsey Domar and Richard Musgrave once noted: "By imposing an income tax on the investor, the treasury appoints itself as his partner" (p. 389). The firm, of course, is not likely to consider the interests of its (silent) partner when making its investment decisions; and its failure to do so must lead to an investment schedule which is (in the view of the government) nonoptimal, since—because of its risk-spreading capabilities—the government's attitude toward an independent risky investment must essentially be one of risk indifference.

That a firm's failure to consider the interests of its tax collector will lead to a nonoptimal level of investment seems virtually certain. One's intuition, perhaps, as well as Mayshar's analysis, leads to the conclusion that firms will invest too *little* in risky projects, unless the government increases the attractiveness of such projects by (for example) subsidizing the firm's cost of capital. But such a conclusion will not in general be correct. Whether a firm will invest more or less heavily in risky projects than would a risk-indifferent government depends critically upon the nature of the uncertainty faced by the firm. Mayshar's conclusion is derived from a model which assumes a particularly simple form of multiplicative uncertainty; as I shall show, it is easy to construct counterexamples in which risk-averse firms *over*invest in risky projects. This point can be made with a model which is considerably less complicated than the one used by Mayshar.

Consider a firm which uses a single input Y (land) to produce a single output C (corn). The number of units of corn produced at the end of the single period being considered is of course an (increasing) function of the number of units of Y employed; but also assume the output level to be a function of a random variable θ (the weather). Thus we have the production function $c = f(y, \theta)$, where y is the number of units of Y invested at the beginning of the period, and c is the number of units of C produced at the end of the period. We assume $\partial f/\partial y > 0$, $\partial^2 f/\partial y^2 < 0$, and $\partial f/\partial \theta > 0$.

It is assumed that the firm has w units of assets at the beginning of the period. If not used to invest in units of Y, these assets may

*Assistant professor, department of economics, Rutgers College. In preparing the final version of this note I have benefited from the comments of an anonymous referee.

be invested in riskless government bonds; thus $y + b = w$, where b is the amount invested in bonds at a safe rate of return r. We also assume that the firm can borrow at the same rate, so that b may be negative.

The firm's total assets at the end of the period are

$$(1) \qquad A \equiv f(y, \theta) + (1 + r)b$$

Let us suppose the firm is subject to a tax at rate t, levied on its "income"—the difference between its end-of-period wealth A and its initial wealth w. Hence the firm's "net worth" at the end of the period is

$$(2) \quad \Pi \equiv A - (A - w)\, t = (1 - t)\,[f(y, \theta) + (1 + r)\, b] + tw$$

Suppose also the firm maximizes the expected utility from wealth $EU(\Pi)$, where E is the expected-value operator and U is a von Neumann-Morgenstern utility function. The utility-maximizing investment y is thus the solution to

$$\partial[EU(\Pi)]/\partial y = E[U'(\Pi)\,(1 - t)\,(\partial f/\partial y - (1 + r))] = 0$$

which requires, if $t < 1$,

$$(3) \qquad E[U'(\Pi)(\partial f/\partial y - 1 - r)] = 0$$

(Dividing (3) by $U'(\Pi)$ yields Mayshar's equation (4).)

Notice from equation (3) that the firm's utility-maximizing investment strategy is completely unaffected by the existence (or the magnitude) of the income tax rate t. But although the firm rationally ignores the existence of the tax, a social planner should not. Let $U_1(T)$ be the planner's utility function, with tax receipts $T \equiv (A - w)t$ as the argument; and let ρ be the true social cost of capital. Define the socially optimal investment strategy y^* as the value of y which maximizes $E[U(\Pi) + U_1(T)]$; the necessary condition for welfare maximization is then

$$(4) \quad \partial\,[E[U(\Pi) + U_1(T)]]/\partial y = E\{(\partial f/\partial y - 1 - \rho)[(1 - t)U'(\Pi) + tU_1'(T)]\} = 0$$

(If the planner's utility function $U_1(T)$ is replaced by a summation of the utility functions of the n consumers who benefit from the tax receipts $\Sigma_{j=1}^{n} \lambda_j E_j U_j(\cdot)$, then (4) is equivalent to Mayshar's equation (8).)

The first question to be considered is whether the value of y which satisfies the firm's first-order condition (3) will also satisfy the "social optimum" (4). An immediate difficulty occurs if the private discount rate r is not the same as the social rate of discount ρ; suppose, however, that the difficulty does not arise, so that $\rho = r$. In that case it is clear that if $t = 0$, then (3) and (4) are identical: if the private cost of capital equals the social cost of capital, and if there is no income tax, then the firm's utility-maximizing investment strategy is socially optimal.

The private and social optima will generally diverge if $t > 0$, however. Suppose $r = \rho$, and let y take the value which would be chosen by a privately owned firm, so that equation (3) is satisified. Then

$$(5) \quad \partial[E[U(\Pi) + U_1T)]]/\partial y = tE[(\partial f/\partial y - 1 - \rho)(U_1'(T) - U'(\Pi))]$$

If both the firm and the planner are risk indifferent, then $U_1'(T)$ and $U'(\Pi)$ are constant numbers, and it is clear that (5) equals zero if (3) is satisfied. Thus risk indifference is a sufficient condition for the firm's investment strategy to satisfy the social optimum.

In the absence of a frictionless capital market, we should not expect firms to be risk neutral, although the economy as a whole is likely to be; that is, although $U_1'(T)$ may be a constant number, $U'(\Pi)$ will be a (decreasing) function of Π. Hence we may write (5) as

$$(5') \quad U_1'(T)\, tE[\partial f/\partial y - 1 - \rho] - tE[U'(\Pi)(\partial f/\partial y - 1 - \rho)]$$

The firm rationally chooses a value of y which equates the second term in (5′) to zero. If $U(\Pi)$ is concave, then the first term will *not* equal zero; but whether it is positive or negative can not be predicted a priori, as we now show. Define

$$B \equiv \partial f/\partial y - 1 - \rho$$

so that if $\rho = r$ the first-order condition (3) may be written as $E[BU'(\Pi)] = 0$. Now $E[BU'(\Pi)] = EB\,EU'(\Pi) + cov(B, U'(\Pi))$. If $cov(B, U'(\Pi)) < 0$, then $EB > 0$ if $E(BU'(\Pi)) = 0$. Thus (5′) is positive if (3) is satisfied. It follows at once that $\partial E[[U(\Pi) + U_1(T)]]/\partial y$ must be positive when (3) is satisifed; and since the left side of (4) is a decreasing function of y, the socially optimal investment y^* must be greater than the private optimum.

This is the case considered by Mayshar. In his analysis $cov(B, U'(\Pi]] < 0$ because of the simplifying assumption that $f(y, \theta) = \theta f(y)$. In that case $B \equiv \theta f'(y) - 1 - \rho$ is clearly an increasing function of θ, while $U'(\Pi)$ is decreasing in θ. But in general $\partial f/\partial y$ need not be increasing in θ. If, for example, $f(y, \theta) \equiv g(y) + \theta$, then $\partial f/\partial y$ is independent of θ, B is nonstochastic, and the social and private optimums are identical regardless of the firm's attitude toward risk.

There is absolutely no reason to rule out a priori the possibility that $\partial f/\partial y$ is decreasing in θ. In that case $cov(B, U'(\Pi)) > 0$, and $EB < 0$ when $E(BU'(\Pi)) = 0$: a risk-averse firm invests *more* in risky projects than is socially optimal.[1] Any government subsidy would drive the firm further from the socially optimal investment strategy.

[1]Suppose, for example, that y is an agricultural firm's investment in water storage facilities, and assume that the random variable θ is rainfall If θ is high, production and profits are high, but the marginal productivity of previously stored water will surely be low Thus $cov(B, U'(\Pi)) > 0$, and the firm overinvests in y. Since the water storage project will increase profits in poor years, risk-averse firms protect themselves against the risk of lower-than-expected profits by overinvesting in water storage. For a similar example, see Kenneth Arrow and Robert Lind, p 377.

Mayshar's analysis is not limited to cases in which the government is risk indifferent; it is sufficient to assume that the government (acting as a proxy for the recipients of the tax revenues) is more nearly risk neutral than is the firm. My analysis is likewise unaffected by an assumption that $U_1(T)$ is concave, but less so than $U(\Pi)$;[2] the critical question is still the relationship between $\partial f/\partial y$ and the random disturbance θ. Unless planners have available some industry-specific information about the (expected) relationship between marginal products and random disturbances, a recommendation that governments subsidize risky private ventures would appear to be premature.

[2]That is, if the Pratt-Arrow coefficient of absolute risk aversion, $\alpha \equiv -U''(\cdot)/U'(\cdot)$, is smaller for the planner's utility function than for the firm's utility function. The method of proof is straightforward; see Duncan Holthausen or the author

REFERENCES

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

E. D. Domar and R. A. Musgrave, "Proportional Income Taxation and Risk Taking," *Quart. J. Econ.*, May 1944, *58*, 388–442.

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

J. Mayshar, "Should Government Subsidize Risky Private Projects?," *Amer. Econ. Rev.*, Mar. 1977, *67*, 20–28.

M. B. Stewart, "Factor Price Uncertainty with Variable Proportions," *Amer. Econ. Rev.*, June 1978, *68*, 468–73.

Should Government Subsidize Risky Private Projects?: Reply

By JORAM MAYSHAR*

The practice of governmental subsidies for private investment projects is widespread in both the developed and the developing nations.[1] Aside from political-distributional reasons, the common economic rationale given for such practices (when no externalities or significant returns to scale are involved) is one of correcting for imperfections in capital markets. Such market imperfections, it is often alleged, result in too little private investment. In their seminal paper Kenneth Arrow and Robert Lind claimed that deficiencies in capital markets do *not* provide such a rationale for subsidization. The purpose of my 1977 paper was to show that if government shared in the yield of a risky investment project, there would be a prima facie case for subsidizing it in order to correct for private underinvestment.

The main contribution of Marion Stewart's paper, as I see it, is in the simple representation of the somewhat cumbersome mathematical representation of the argument in my paper, which he achieved by endowing the "social planner" with preferences of his own. Stewart however also challenges the generality of the case for subsidization and provides a counter example. He further suggests that because of difficulties in obtaining the necessary information the general recommendation for subsidization may be unwarranted. I would like here both to explain why I consider his caveat to be of minor importance and to comment somewhat more generally on the applicability of the argument for subsidization.

In Stewart's counterexample the output generated by a marginal investment in a project $f_y(y, \theta)$ is negatively correlated to that of the total output $f(y, \theta)$, and thus also to the bulk of the output of intramarginal or initial investment. In this case the marginal investment serves as a form of self-insurance for the private firm. If (because of independence) the government is risk neutral towards the risks of its share in the project's output, any resources spent on such self-insurance must be considered a waste from a social point of view; marginal investment ought thus to be discouraged.

While the form of the production function $f(y, \theta)$ is clearly more general than the multiplicative form $f(y)\phi(\theta)$ which I used, Stewart's example is, I believe a theoretical anomaly with little practical significance. A more revealing generalization of the multiplicative case, in which not only the scale of investment but the choice of technique can be depicted as well, was suggested in footnote 4 in my paper. Accordingly, let a firm be able to produce a risky output $x_0 + \Sigma_{k=l}^{K} x_k \phi_k(\theta)$ by a choice of a technique $x = (x_0, x_1, \ldots, x_k)$ whose total costs are $y = C(x)$. For each level of costs y, the firm may be thought of as choosing the optimal technique $x(y)$, and this will define in turn output as a function of the investment level $f(y, \theta)$, as used by Stewart. While some of the techniques might be utilized to reduce the risk of production, it is likely that such self-insuring techniques would be used whatever the level of investment. Still it is to be expected that self-insurance will rarely dominate, and an increase in the overall scale of investment will increase the overall risks as well.

Stewart's suggestion of viewing the investment in a water storage facility by an agricultural firm as a counterexample considers that particular project as being somehow divorced from the other activities of that firm. It is likely however that the decision concerning the scale of such a self-insuring technique will be made simultaneously with decisions on the

*Hebrew University.

[1]For a wide interpretation and survey of the practice of subsidization see for example the article by Alan Prest and the volume in which it appears.

scale of the land and machinery to be used, the choice of crops and a multitude of other technique decisions. If the interdependence of all these investment decisions is recognized, it is quite likely that overall risks will increase with the scale of operation. Only a modest amount of information may be needed to ascertain that this is indeed the case. For the government however to involve itself in the particulars of how the firm allocates its investment funds and to try to discourage the building of the water storage facility may indeed be a "premature" recommendation, but this was clearly not the intent of my paper.

The argument for subsidization rests on the government's role in complementing (or even substituting) the imperfect capital market by spreading to taxpayers the risks of its share in private firms' profits. A further informational difficulty is the need to estimate taxpayers' own evaluation of these risks. Given these informational requirements, what I perceive as feasible, and in general desirable, would be very broad and simple investment subsidization schemes. These can be differentiated on an industry basis, possibly according to each industry's risk and correlation with the business cycle, and possibly also differentiated according to the degree of risk spreading achieved by each firm via the market.

Indeed the subsidization programs adopted in the United States which take such forms as investment tax credits, accelerated depreciation allowances, loan guarantees, and the like can often be rationalized by the above considerations.[2] Incentives for oil explorations, for example, may be explained both by the large risks involved and by the consideration that the fortunes of that industry may be largely uncorrelated or even negatively correlated (because of the effects of the international price of oil) to those of the economy. On the other hand, subsidies to farmers and to new and small businesses in general may be rationalized in that these are cases where the deficiency of the market in spreading risks is the most obvious. In fact, the same reasoning will also apply to investment conducted by individuals directly. Thus if investment in education were risky (as it seems to be),[3] the government sharing in the risky yield to education through the income tax would provide an argument for subsidizing it.

While taxes on profits and on income provide the mechanisms in which most developed Western governments share in the risky yield to private investments, other forms for such taxation are also possible. Thus a nationalized marketing board for the export of an agricultural or a mining product, as is prevalent in many developing countries, usually serves a similar purpose of expropriating some of the risky profits of private production.

The most extreme example of countries in which the government substitutes for the capital market and provides the only means for the spread of risks is the communist countries. To illustrate the applicability of the same argument to such economies, the cultivation of the virgin lands in the Soviet Union in the 1950's may be cited.[4] The harsh and unreliable conditions in that region explain why these lands were not cultivated, previous to the government involvement, even though weather conditions there are apparently negatively correlated to those in the European grain districts. This negative correlation may imply however the social desirability of the cultivation of such lands if the risks could be shared by all. Capital markets would have to be almost inconceivably "perfect" to obtain such an outcome by unassisted market forces. In the absence of such capital markets, the state may have provided in this case a sufficient substitute to the pooling and spreading of risks to justify the undertaking of that risky venture. The gamble apparently paid off in

[2]Note however that exempting a fraction of profits from taxes as an incentive would undermine the basis which was suggested here for subsidization.

[3]A case which initially I thought might have provided an example for Stewart's case of self-insurance is that of graduate level education. Empirical results however indicate that even in this case the risks (measured by a cross-section variance) increase with the investment level.

[4]The facts concerning this example are based on Theodore Shabad's article in the *New York Times*; the possibly far-fetched interpretations are of course my own.

1973 when the virgin lands provided record crops which softened the impact of the serious crop failures in the Ukraine.

REFERENCES

K. J. Arrow and R. C. Lind, "Uncertainty and the Evaluation of Public Investment Decisions," *Amer. Econ. Rev.*, June 1970, *60*, 364–78.

J. Mayshar, "Should Government Subsidize Risky Private Projects?," *Amer. Econ. Rev.*, Mar. 1977, *67*, 20–28.

A. R. Prest, "The Economic Rationale of Subsidies to Industry," in Alan Whiting, ed., *The Economics of Industrial Subsidies*, London 1976.

T. Shabad, "Soviet Reports Best Asian Crop: Khruschev Gamble to Farm in East Apparently Pays," *New York Times*, Oct. 15, 1972.

M. B. Stewart, "Should Government Subsidize Risky Private Projects?: Comment," *Amer. Econ. Rev.*, June 1979, *69*, 459–461.

Consumer's Surplus Without Apology: Comment

By GEORGE W. MCKENZIE*

Over the past several years there has been a considerable resurgence of interest in the theory of consumer's surplus with the general aim of identifying those circumstances under which it may serve as a viable welfare indicator. In a recent issue of this *Review* and in this vein, Robert Willig has attempted to establish criteria or rules of thumb for determining when consumer's surplus may reasonably approximate the equivalent and compensating variations. Unfortunately, his striking conclusions are based on some rather special assumptions which severely limit the applicability of his approach.

In this comment, I shall first examine the two main problems with Willig's approach: (a) it cannot be easily generalized to cases where the prices of several commodities vary; (b) it is valid only when price and income changes are relatively small. Second, I shall present some numerical examples to illustrate that the magnitudes of error involved in Willig's approach may be far from negligible. Finally, his procedures are contrasted with the one recently presented by Ivor F. Pearce and the author, which enables an *exact*, not an approximate, money metric welfare indicator to be constructed on the basis of observable information.

I. Two Errors

Most of the projects and policies to which consumer's surplus techniques will be applied will involve variations in income and several prices. However, Willig's approach is incapable of such *general* application. There are two reasons for this. The first difficulty arises from his attempt to construct a welfare measure based on a constant income elasticity of demand (pp. 592–93). Following his notation, let m be income and p_i and X^i be the price and quantity demanded of commodity i. If the constant income elasticity of demand for this good is denoted as

$$n_i \equiv \frac{\partial X^i(p,m)}{\partial m}\frac{m}{X^i(p,m)}$$

then, following Willig, it is possible to show that

$$X^i(p,m^2) = X^i(p,m^1)\left(\frac{m^2}{m^1}\right)^{n_i} \qquad (1)$$

where m^2 and m^1 are two different income levels and n_i is the *constant* income elasticity of demand associated with the ith commodity. However, it is necessary to note that if we differentiate the budget constraint $m = \Sigma p_i X^i$ with respect to income and rearrange variables we obtain

$$\sum \alpha_i n_i \equiv 1 \qquad (2)$$

where α_i is the value share of commodity i in total income, $p_i X^i/m$. In general, if all n_i are to be constant, then all α_i must also be constant. This condition unambiguously implies that consumer preferences are homothetic, that is, all income elasticities equal one.

But the restriction must arise in another way as well. Consider the consumer's surplus measure utilized by Willig:

$$A \equiv \int_{p_i^1}^{p_i^2} X^i(p,m^0)\,dp_i \qquad (3)$$

As Harold Hotelling showed in his classic article, it is possible to generalize this measure to the case of more than one good only if the matrix of uncompensated price effects is symmetric. In turn this condition implies homotheticity.[1] Although Willig does argue that "measured income elasticities of

*Senior lecturer in economics, University of Southampton, England. This paper is based on research sponsored under a project grant from the *SSRC* (UK). I would also like to thank Ivor Pearce and Peter Wagstaff for helpful comments and encouragement.

[1]On this point the reader should also consult Paul Samuelson and Eugene Silberberg.

demand tend to cluster closely about 1.0, with only rare outliers" (p. 590), he does not present any empirical evidence to support this statement. It is one with which most economists would probably disagree.[2]

Willig's second error arises in his attempt to relate the compensating and equivalent variations to his consumer's surplus measure A. Following Willig let us write his compensation function as

$$(4) \qquad C = m^1\left[1 + \left[\frac{1-n_i}{m^1}\right] A\right]^{1/(1-n_i)}$$

He then attempts to "loosely" approximate (4) via a second-order Mclaurin series with respect to A to obtain

$$(5) \qquad CV \approx A + \frac{n_i A^2}{2m^1}$$

$$(6) \qquad EV \approx A - \frac{n_i A^2}{2m^1}$$

where CV and EV are the compensating and equivalent variations, respectively. However, such an approximation procedure neglects the fact that the remainder term involved may be quite substantial particularly if A is not very small.

This breakdown in Willig's approach should not really come as any great surprise. The ability of any function to approximate another function is determined by three factors: 1) the initial values of relevant variables; 2) the magnitude of any changes in these variables; and 3) the parameters of the function being approximated. Thus even if we restrict the range of values assumed by prices and income, as Willig does, we still need information about the parameters determining consumer preferences before we can say anything about the accuracy of an approximation. And this is true even when preferences are homothetic.

II. A Numerical Example

To illustrate this point, let us assume that consumer preferences can be described by the utility function developed by Lawrence Klein and Herman Rubin and by Richard Stone:

$$(7) \qquad U = \sum b_i \, ln \, (X_i - c_i)$$

where b_i is the marginal propensity to consume the ith commodity and the c_i are parameters associated with the ordinary price elasticities. This may be determined by examining the demand functions obtained by maximizing (7), subject to the budget constraint:

$$(8) \quad X^i = (1 - b_i)c_i + \frac{b_i m}{p_i} - b_i \sum_{j \neq i} \frac{p_j c_j}{p_i}$$

Following Stone and Giovanna Croft-Murray, (pp. 75–76), we may write the compensating and equivalent variations associated with (7) as follows:

$$(9) \quad CV = (m^1 - \sum p_i^1 c_i) \prod_j (p_j^2/p_j^1)^{b_j} - (m^2 - \sum p_i^2 c_i)$$

and

$$(10) \quad EV = (m^2 - \sum p_i^2 c_i) \prod_j (p_j^1/p_j^2)^{b_j} - (m^1 - \sum p_i^1 c_i)$$

In addition, we may calculate from (8), the sum of areas under the ordinary demand functions as

$$(11) \quad A = \Delta m - \sum\left[(1 - b_i)c_i p_i + b_i(m^1 - \sum_{j \neq i} c_j p_j^1) \, ln \, p_i\right]\Big|_{p_i^1}^{p_i^2}$$

For simplicity, it is assumed that there are only two commodities, and that both of their prices and total income vary. The initial variable and parameter values are as follows:

$$m^1 = 3000 \qquad b_1 = .3 \quad c_1 = -600$$
$$p_1^1 = p_2^1 = 1 \quad b_2 = .7 \quad c_2 = -200$$

All commodities are therefore gross substitutes. Now let us compare this initial situation (called Case 1) with the three cases shown in Table 1.

As is well known, the equivalent variation is a true ordinal welfare indicator[3] and on the

[2]For example, see the paper by Laurits Christensen, Dale Jorgenson, and Lawrence Lau who reject restrictions implied by homotheticity under a number of conditions.

[3]For example, see Willig, p. 591. The differences in magnitude between EV and ΔU in Table 1 are due simply to the scaling effect involved in the monotonic transformation which links them.

TABLE 1[a]

Case	p_1^t	p_2^t	m^t	CV	EV	ΔU	A	$\frac{A-EV}{EV}$	$\frac{A-CV}{CV}$
2	1	1	4000	1000	1000	.02361	1000	0	0
3	1.5	.5	2337	695	1000	.02361	874	−.126	+.205
4	.5	1.5	4578	1079	1000	.02361	1042	+.042	−.037

[a]Numbers may not be exactly consistent due to rounding.

basis of this it can be seen that the three alternative cases, 2–4, all lie on the *same* indifference surface which is preferred to the initial situation. However, the consumer's surplus indicator A is incapable of identifying this result: Case 3 is shown to be 12.6 percent worse than Case 1 whereas Case 4 is calculated as being 4.2 percent better than Case 1. Whereas Willig's approach would be reasonable for variations in the price of a single commodity, it is simply incapable of generalization to more realistic situations where several prices and income are affected by the introduction of a project or policy. It is not possible to add changes in income to the consumer's surplus measure A, the procedure suggested by Willig, to obtain an accurate welfare indicator.

III. A Correct Alternative

Both EV and U in the previous exercise are cardinal representations of the assumed ordinal preference function. EV, however, is the only such representation with a meaningful interpretation in terms of monetary units. As such, this measure must form the basis for cost-benefit analysis, not consumer's surplus. In contrast, Willig claims that EV is not expressable in terms of observable units (p. 589). This is not the case, however. As Pearce and I have shown elsewhere,[4] it is possible to write a welfare indicator which possesses the following three properties: 1) It is capable of correctly ranking all relevant, alternative price/quantity situations for an individual or homogeneous group of individuals. 2) It is expressible in terms of monetary units. 3) It is capable of expression in terms of parameters of ordinary, observable demand functions.

Our proposed welfare indicator thus has those desirable properties of consumer's surplus, 2) and 3), but it is also an *exact* not an approximate welfare indicator. Further it is equal to the equivalent variation.

To summarize the argument given by Pearce and myself, note first that we may write any utility function

$$U = U(X^1, \ldots, X^n) \tag{12}$$

as an indirect cost of utility function[5]

$$e(U) = g(m, p, p_1^t, \ldots, p_n^t) \tag{13}$$

Since e and m are both measured in terms of the same monetary units, the marginal utility of money (λ) equals one, given *initial* prices, and all its higher derivatives with respect to income $\partial^r\lambda/\partial Y^r$ equal zero, again given initial prices.

These results regarding λ then enable us to express any utility function in terms of a Taylor-series expansion involving only the parameters of ordinary demand functions, and eliminating all traces of the marginal utility of money and its derivatives with respect to prices and income. For example, in circumstances where a second-order approximation is valid, the McKenzie-Pearce welfare indicator is written as

$$\Delta e = -\sum X_i \Delta p_i - \frac{1}{2}\sum_i \sum_j \left(\frac{\partial X_i}{\partial p_j} - X_i \frac{\partial X_j}{\partial m}\right)\Delta p_j \Delta p_i + \Delta Y - \sum \frac{\partial X_i}{\partial Y}\Delta p_i \Delta m \tag{14}$$

[4]For a fuller discussion, the reader should consult our paper. Also relevant are my 1976 and 1977 papers.

[5]Also see Willig, equation (10), p. 591 and fn. 16, p. 596.

Higher-order terms may be added so as to achieve an approximation to *any* desired degree of accuracy. Indeed in the numerical examples given above this action would be a *necessity*. As additional terms are added, the change indicated by Δe will approach the true equivalent variation in the limit (see my 1976 paper). Such a procedure is precluded by the approach taken by Willig.

While the approach suggested by Pearce and myself involves fairly tedious calculations, this is only to be expected of a solution to a complex problem. The key thing to note in comparing our approach with that of Willig's is that both require the same basic information: knowledge of observable consumer demand functions. On the other hand, such data are necessary to perform the integration indicated by equation (3). They are also required to calculate the Taylor-series approximations of the sort indicated by equation (14). But since the latter approach can always generate a more accurate welfare indicator than consumer's surplus, it is the one which should be used in practice.

To sum up the argument of this comment, there is a strong case for rejecting Willig's conclusion that "cost-benefit welfare analysis can be performed rigorously and unapologetically by means of consumer's surplus" (p. 596), since he has been unable to establish *general* conditions under which consumer's surplus is either theoretically sound or empirically accurate. Instead it is possible to create a logical basis for cost-benefit analysis involving the procedures which Pearce and myself have discussed elsewhere. The argument that there is no other applied welfare indicator but consumer's surplus is thus no longer a valid one.

REFERENCES

L. R. Christensen, D. W. Jorgenson, and L. D. Lat "Transcendental Logarithmic Utilit Functions," *Amer. Econ. Rev.*, June 197 *65*, 367–83.

H. Hotelling, "The General Welfare in Relation to Problems of Taxation and of Railway and Utility Rates," *Econometrica* July 1938, *6*, 242–69.

L. R. Klein and H. Rubin, "A Constant Utilit Index of the Cost of Living," *Rev. Econ Stud.*, No. 2, 1948, *15*, 84–7.

G. McKenzie, "Measuring Gains and Losses, *J. Polit. Econ.*, June 1976, *84*, 641–46

———, "Complementarity, Substitutabilit and Independence," *Oxford Econ. Pap.* Nov. 1977, *29*, 430–41.

——— **and I. F. Pearce,** "A New Operationa Procedure for Evaluating Economic Policies," *Rev. Econ. Stud.*, Oct. 1976, *43* 465–68.

P. Samuelson, "The Constance of the Margina Utility of Income," in Oskar Lange et al. eds., *Studies in Mathematical Economic and Econometrics in Memory of Henr Schultz*, Chicago 1942; reprinted in *Th Collected Scientific Papers of Paul A Samuelson*, Cambridge 1965.

E. Silberberg, "Duality and the Many Consumer's Surpluses," *Amer. Econ. Rev.*, Dec 1972, *62*, 942–52.

Richard Stone, "Linear Expenditure System and Demand Analysis: An Application to the Pattern of British Demand," *Econ. J.* Sept. 1954, *64*, 511–27.

——— **and G. Croft-Murray,** *Social Accounting and Economic Models*, London 1959.

R. D. Willig, "Consumer's Surplus Withou Apology," *Amer. Econ. Rev.*, Sept. 1976 *66*, 589–97.

Consumer's Surplus Without Apology: Reply

By ROBERT D. WILLIG*

I began my article "Consumer's Surplus Vithout Apology" (henceforth, *CSWA*) with he words "The purpose of this paper is to ettle the controversy surrounding consumer's urplus . . . " (p. 589). That is my purpose ıere, as well. However, George McKenzie's ublished comments have taught me that if ırticles and careful analyses settle controveries, they only do so very slowly.

McKenzie makes sweeping and attacking tatements about *CSWA* but fails to substaniate them. He scatters birdshot criticisms at *'SWA* that are based on misreadings of ather clear material. He offers an example in vhich he miscalculates multiproduct consumr's surplus. Finally, he contends that his own with Ivor F. Pearce) approach to welfare nalysis is preferable to the consumer's urplus approach.

In this reply, to keep the record straight, I how in Section I that each of McKenzie's trongly worded attacks is unsubstantiated nd invalid, and that each of his more technial sounding criticisms rests only on misreadngs of *CSWA*. More interestingly, in Section I, I summarize some of the theory of multiroduct consumer's surplus that is needed to nderstand the calculation error in and roper interpretation of the example that AcKenzie proffers. In Section III, I argue hat the approach to welfare analysis advoated by McKenzie and Pearce is far less ıseful than the consumer's surplus methodlogy.

I. McKenzie's Misreadings and Unfounded Attacks

McKenzie states that he has made "a trong case for rejecting Willig's conclusion hat 'cost-benefit welfare analyses can be performed rigorously and unapologetically by means of consumer's surplus', . . ." (p. 468). What is this "strong case"? Apparently, it is summarized by "Unfortunately, his [Willig's] striking conclusions are based on some rather special assumptions which severely limit the applicability of his approach" (p. 465). McKenzie does not proceed to delineate the "rather special assumptions" that he has in mind. Instead, he lists "two main problems with Willig's approach: (a) it cannot be easily generalized to cases where the prices of several commodities vary; (b) it is valid only when price and income changes are relatively small" (p. 465).

It is striking that McKenzie never deals with my approach to these areas. Footnote 2 in *CSWA* reads "While I restrict attention to single price changes here, analogous, but more complex formulae are derived for multiple price changes in my papers (1973a,b)" (p. 589). These papers show that (a) above is false. Some of the pertinent results from these papers are summarized in Section II, below.

With regard to "problem" (b), it is difficult to infer what McKenzie has in mind. For example, the bounds derived in *CSWA* can be small for price and income changes that are large relative to the initial levels of these parameters. I am unable to respond further to McKenzie on "problem" (b) since, after listing it, he does not mention it again.

McKenzie entitles his Section I "Two Errors." The first "error" he accuses me of making is having undertaken the analysis of the case in which there is a constant income elasticity of demand. He bases his disapproval on the very well-known fact that if the income elasticity of demand for each good were constant, then those elasticities would all be one,[1] and that this would be an unrealistic

*Professor of economics and public affairs, Princeton Jniversity. Part of this paper was written while I was a ember of the economics research department of Bell aboratories. I am grateful to Magina Jack and Edward ajac for their editorial suggestions.

[1]McKenzie states that if the income elasticities of demand of all goods are constant then the value shares of all goods must also be constant. This is false. The budget constrained maximization of any homothetic utility function yields demands that all have constant unitary income

situation.[2]

However, in *CSWA* I did not base my argument for consumer's surplus on the case of constant income elasticities. Instead I presented it as an instructive special case that introduced some of the techniques utilized to analyze in Section IV the general case of nonconstant income elasticity of demand. The concluding sentence of Section III reads "The next section will establish this formula rigorously for nonconstant income elasticity of demand" (1976a, p. 593).

In McKenzie's view, my "second error" was "loosely" applying (on p. 593) second-order approximations to the derived exact expressions for the compensating and equivalent variations. He felt the need to remind the reader that "such an approximation procedure neglects the fact that the remainder term involved may be quite substantial particularly if A is not very small" (p. 466).[3]

This view of *CSWA* reflects another surprisingly distorted misreading. It was I who used the world "loosely" (p. 593) to describe applying second-order approximations for only heuristic purposes. The entire Section IV of *CSWA* was devoted to calculating, analyzing, and bounding the remainder term in question. Indeed, the principal contribution was the derivation of "precise upper and lower bounds on the percentage errors of approximating the compensating and equivalent variations with consumer's surplus" (1976a, p. 589).

elasticities. However, only Cobb-Douglas utility yields demands with constant value shares The restrictions imposed on integrable demand functions by locally constant elasticities are studied in my 1976b paper.

[2]McKenzie seems to think that I believe in the realism of homothetic preferences because I wrote "Measured income elasticities of demand tend to cluster closely about 1.0, with only rare outliers" (1976a p 590) Homotheticity was not at issue. Instead, as clearly stated in the next sentence, I was arguing that one-half the product of A/m and the income elasticity of demand would typically be small if A/m were small. This argument does not require measured income elasticities to cluster nearly as closely about one as would a belief in homothetic preferences.

[3]Ironically, it is the McKenzie and Pearce approach, rather than *CSWA*, that neglects remainder terms from finite Taylor-series expansions. See Section III, below.

Another misstatement by McKenzie that must not be permitted to stand is "It is not possible to add changes in income to the consumer's surplus measure A, the procedure suggested by Willig, to obtain an accurate welfare indicator" (p. 467). To the contrary, as established in *CSWA*, change in income plus consumer's surplus *is* a measure that can be *rigorously* used in welfare analysis. The relationships in (4) on page 590 say that the consumer finds changes in prices and income beneficial (injurious) if and only if the income change minus the compensating variation is positive (negative). Similarly, footnote 16 says that income m plus the equivalent variation, $\mu(p^0|p, m) - \mu(p^0|p^0, m)$, accurately ranks the welfare desirabilities of all combinations of prices and income.

The principal point of *CSWA* is that in most practical applications consumer's surplus can be substituted with little loss of accuracy for the compensating and equivalent variations in these rigorous welfare calculations. The size of the numerical error caused by this substitution can be assessed from observables by means of the bounds presented in *CSWA*. Thus, as symbolized in (24) or in analogous expressions using the equivalent variation, change in income plus consumer's surplus is a measure that can be utilized to perform individual welfare analysis both rigorously and unapologetically.

II. Multiproduct Consumer's Surplus

In this section, I discuss McKenzie's numerical example and summarize some of the theory of multiproduct consumer's surplus that is required to properly interpret it.

As noted by McKenzie, it is well known that for multiple price changes, the numerical value of the consumer's surplus line integral,[4]

$$(1) \qquad \oint_\Gamma \sum_i X^i(p, m^0)\, dp_i$$

where Γ is some path between the initial prices p^0 and the final prices p', generally

[4]Here, and below, I follow the notation used in *CSWA*. In particular, $X^i(p, m)$ is the demand for good i at prices p and income m.

depends on the path of integration Γ. Consequently, various absurdities can be generated by clever choices of the path.

For practical purposes, the typical choice is a rectangular path along which the prices move from their initial to their final values one at a time, while those that have already moved stay at their new values. Figure 1 illustrates the two rectangular paths for the case of changes in two prices. The first path is comprised of the directed line segment running from (p_1^0, p_2^0) to (p_1', p_2^0), followed by that running from (p_1', p_2^0) to (p_1', p_2'). Along the first segment, only p_1 changes, while p_2 is held at its initial value p_2^0. In contrast, along the second segment, p_1 is held at its new value p_1', while p_2 changes from p_2^0 to p_2'. Thus for this path, (1) becomes

$$(2)\quad \int_{p_1^0}^{p_1'} X^1(p_1, p_2^0, p_3^0, \ldots, p_n^0, m^0)\, dp_1 + \int_{p_2^0}^{p_2'} X^2(p_1', p_2, p_3^0, \ldots, p_n^0, m^0)\, dp_2$$

Obversely, the second path has p_2 changing with the $p_1 = p_1^0$, and p_1 moving from p_1^0 to p_1' while p_2 is held at its new value p_2'. For this path, (1) equals

$$(3)\quad \int_{p_1^0}^{p_1'} X^1(p_1, p_2', p_3^0, \ldots, p_n^0, m^0)\, dp_1 + \int_{p_2^0}^{p_2'} X^2(p_1^0, p_2, p_3^0, \ldots, p_n^0, m^0)\, dp_2$$

While these constructions may seem evident, they were not utilized by McKenzie. Instead, in his equation (11), he calculates his version of multiproduct consumer's surplus by integrating along the directed segments running (in the present notation) from (p_1^0, p_2^0) to (p_1', p_2^0) and running from (p_1^0, p_2^0) to (p_1^0, p_2'). As is clear from Figure 1, these segments do not form a path from (p_1^0, p_2^0) to (p_1', p_2'). Of course, if X^1 were insensitive to p_2 and X^2 were insensitive to p_2, then McKenzie's calculation would agree with both (2) and (3). However, with nonzero cross-elasticities of demand (as is the case in McKenzie's (11)), McKenzie's procedure is incorrect, and it would yield wrong answers even if income effects were negligible.

The consumer's surplus line integral (1) can be related to the compensating and equivalent variations via the same sequence of steps outlined in *CSWA*, Section II. Thus using the same notation,

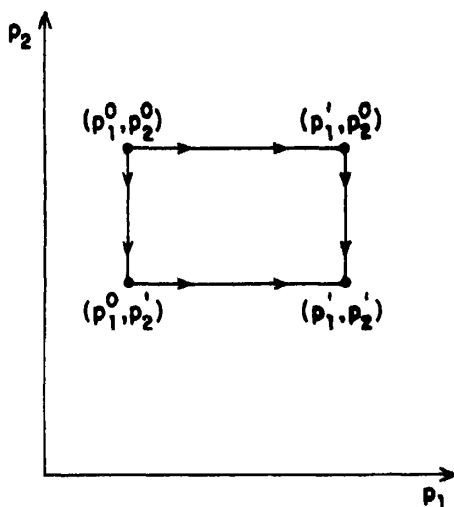



FIGURE 1

$$(4)\quad C = \oint_{\Gamma} \sum_i X^i(p, \mu(p \mid p^0, m^0))\, dp_i$$

$$(5)\quad E = \oint_{\Gamma} \sum_i X^i(p, \mu(p \mid p', m^0))\, dp_i$$

Here, Γ can be any path between p^0 and p' because these line integrals are path independent. This follows from the fact that the integrands are compensated demand functions which are the gradients of the μ functions.

Then, regardless of the path dependence of (1), a particular path Γ can be chosen, and answers can be sought to the questions of how closely consumer's surplus calculated along Γ approximates the true welfare measures C and E.

Let us restrict attention to the rectangular path along which first p_1 changes from p_1^0 to p_1', then p_2 changes from p_2^0 to p_2', etc. Denote

$$A_i = \int_{p_i^0}^{p_i'} X^i(p_1', \ldots, p_{i-1}', p_i, p_{i+1}^0, \ldots, p_n^0, m^0)\, dp_i, \quad A = \sum_{i=1}^{n} A_i$$

Thus, A_i is the consumer's surplus calculated along the ith leg of this rectangular path, and A is the total consumer's surplus along the path. Let $\bar{\eta}_i$ and $\underline{\eta}_i$ be upper and lower bounds, respectively, on the income elasticity of demand for the ith good, over the relevant region.[5]

It has been shown that if all price movements are in the same direction (either $p' \geq p^0$ or $p' \leq p^0$), then multiproduct consumer's surplus bears the same relationships to the compensating and equivalent variations as does consumer's surplus over a single price change. In these cases, the rules of thumb derived in *CSWA*, inequalities (1) and (2), hold for multiproduct consumer's surplus.

THEOREM 1:[6] *Let* $A = \Sigma A_i$, $\bar{\eta} = max \{\bar{\eta}_i | p'_i \neq p^0_i\}$, *and* $\underline{\eta} = min \{\underline{\eta}_i | p'_i \neq p^0_i\}$. *If* $p' \geq p^0$ *or* $p' \leq p^0$, $|\bar{\eta}A/2m^0| \leq .05$, $\underline{\eta} > 0$, *and* $|A/m^0| \leq .5$, *then, to within* .005,

$$(6) \qquad \frac{\underline{\eta}|A|}{2m^0} \leq \frac{C - A}{|A|} \leq \frac{\bar{\eta}|A|}{2m^0}$$

and

$$(7) \qquad \frac{\underline{\eta}|A|}{2m^0} \leq \frac{A - E}{|A|} \leq \frac{\bar{\eta}|A|}{2m^0}$$

The case of price increases mixed with price decreases is far more complex. The primary new difficulty is that A/m^0 may be small in absolute value while its components (for example, A_1/m^0 and A_2/m^0), of opposite signs, have large absolute values. In such a situation, compensated demands can stray significantly far from Marshallian demands along the path of integration, and consequently A can fail to closely approximate C and E. On the other hand, if each of the price changes causes only a moderate change in real income, then relationships analogous to (6) and (7) hold, and A will be relatively close to C and E.

Let $p'_i > p^0_i$ for $i = 1, 2, \ldots, K$; $p'_i < p^0_i$ for $i = K + 1, \ldots, L$; $A_u = \Sigma^K_{i=1} A_i$; $A_d = \Sigma^L_{i=K+1} A_i$; $\bar{\eta}_u = \max \{\bar{\eta}_1, \ldots, \bar{\eta}_K\}$; $\underline{\eta}_u = \min \{\underline{\eta}_1, \ldots, \underline{\eta}_K\}$; $\bar{\eta}_d = \max \{\bar{\eta}_{K+1}, \ldots, \bar{\eta}_L\}$; and $\underline{\eta}_d = \min \{\underline{\eta}_{K+1}, \ldots, \underline{\eta}_L\}$. Note that the subscripts u and d stand for "prices up" and "prices down," respectively.

LEMMA:[7] *Let* $\bar{\eta}_u, \underline{\eta}_u, \bar{\eta}_d$, *and* $\underline{\eta}_d$ *be unequal to* 1, *let all expressions in square brackets, below, be nonegative, and let* $\underline{\eta}_u \geq 0$, $\underline{\eta}_d \geq 0$. *Then*

$$\mu(p'|p^0, m^0) \geq v\left[1 + (1 - \underline{\eta}_d)\frac{A_d}{v}(w/m^0)^{\bar{\eta}_d}\right]^{1/(1-\underline{\eta}_d)}$$

$$\mu(p'|p^0, m^0) \leq w\left[1 + (1 - \bar{\eta}_d)\frac{A_d}{w}\left(\frac{v}{m^0}\right)^{\underline{\eta}_d}\right]^{1/(1-\bar{\eta}_d)}$$

where

$$w = m^0\left[1 + (1 - \bar{\eta}_u)\frac{A_u}{m^0}\right]^{1/(1-\bar{\eta}_u)}$$

$$v = m^0\left[1 + (1 - \underline{\eta}_u)\frac{A_u}{m^0}\right]^{1/(1-\underline{\eta}_u)}$$

This lemma is of independent interest because it provides bounds on C or E that can be calculated from the directly observable surplus measures and income elasticities of demand. However, these bounds are in forms that yield no qualitative information. Detailed numerical analysis of them yields this more readily interpretable result.

THEOREM 2:[8] *Suppose that* $\underline{\eta}_u \geq 0$, $\underline{\eta}_d \geq 0$, $\bar{\eta}_u \leq 10$, $\bar{\eta}_d \leq 10$, $|A_u|/m^0 \leq .25$; $|A_d|/m^0 \leq .25$, $|\bar{\eta}_u A_u/2m^0| \leq .05$, $|\bar{\eta}_d A_d/2m^0| \leq .05$, *and* $|\bar{\eta}_d A_u/2m^0| \leq .05$. *Then, with* $A = A_u + A_d$,

$$(8) \quad C - A \geq \frac{\underline{\eta}_u A_u^2}{2m^0} + \frac{\underline{\eta}_d A_d^2}{2m^0} + \frac{\bar{\eta}_d A_u A_d}{m^0} \cdot \left(1 + \frac{\bar{\eta}_u A_u}{2m^0}\right) + .01A_d - .005A_u$$

[5]See p. 40 of my 1973b paper for the definition of this region in price-income space.

[6]See pp. 40–42 of my 1973b paper for details and proofs.

[7]The proof parallels the reasoning detailed in pp 44–48 of my 1973b paper. Similar results can be established with the same methods for the case in which the price decreases precede the price increases along the path of integration.

[8]An analogous result for bounding A–E can be derived

$$(9)\ C - A \leq \frac{\bar{\eta}_u A_u^2}{2m^0} + \frac{\bar{\eta}_d A_d^2}{2m^0} + \frac{\underline{\eta}_d A_u A_d}{m^0} \cdot \left(1 + \frac{\underline{\eta}_u A_u}{2m^0}\right) - .015A_d + .005A_u$$

Note that (8) and (9) are essentially generalizations of the inequalities in (6). This can be seen by setting either A_u or A_d equal to 0, and dividing through by the absolute value of the other surplus. The terms $.01A_d$, $.005A_u$, and $.015A_d$ are bounds on the errors caused by approximating the functional forms in the lemma. These errors would be potentially larger if the quantitative hypotheses of the theorem were violated. The interaction term involving A_uA_d arises from the view that the compensated income along the portion of the integration path with price decreases begins from approximately $m^0 + A_u + \bar{\eta}_u A_u^2/2m^0$. In contrast, the compensated income is viewed as beginning at m^0 in the derivation of (6).

The qualitatively fruitful interpretation of Theorem 2 is that the error of approximating C with A is bounded between fractions of the absolute values of A_u and A_d. These fractions are small if $|\bar{\eta}_u A_u/2m^0|$ and $|\bar{\eta}_d A_d/2m^0|$ are small. The fractions are not applied to the net surplus A, but instead to the absolute value of each portion. Thus, multiproduct consumer's surplus is only assured to be a close approximation to the compensating (and equivalent) variation if each of the price changes has only a moderate effect on the welfare of the consumer.

From this perspective, the meaning of McKenzie's numerical example is clear. For instance, consider the move from Case 1 in the text to Case 3 in Table 1. This involves halving the price of good 2, and the budget share of good 2 at the higher price is .8. This is a delightfully revolutionary change for the consumer! Proper calculation of the surpluses, along the type of path that underlies Theorem 2, gives $|A_d/m^0| = .59$ and $|\underline{\eta}_d A_d/2m^0| = .24$. No analyst of consumer's surplus has ever claimed that it is generally (with demands realistically sensitive to income) an accurate welfare measure of such revolutionary changes. Instead, it is generally claimed that consumer's surplus is accurate enough for most practical purposes. This claim was rigorously defined and established in *CSWA*, and is reaffirmed here for multiple price changes.

III. Infinite Series Expansions vs. Error Bounds

In his Section III, McKenzie contrasts consumer's surplus with an approach to welfare analysis developed elsewhere by McKenzie and Pearce. This approach utilizes as a welfare measure the Taylor-series expansion of the equivalent variation function. McKenzie states that it is "an *exact* not an approximate welfare indicator" (p. 467), and that since it "can always generate a more accurate welfare indicator than consumer's surplus, it is the one which should be used in practice" (p. 468).

Of course, knowledge of the equivalent variation function would obviate approximating it with consumer's surplus or with any method. Also, it is well known that the limit, as the number of terms goes to infinity, of the Taylor-series expansion of a well-behaved function is the function itself. However, these two facts do not imply that the McKenzie-Pearce approach has significant virtue.

How many terms of the infinite Taylor series would they recommend including in their "approximation"? Perhaps they would continue adding terms until the terms become small, or until the partial sums seem to reveal their limit. These are dangerous procedures that can result in arbitrarily large errors.[9]

I would feel far more comfortable with McKenzie-Pearce if they had focused on the remainder term given by Taylor's Theorem (see Tom Apostol, p. 124). When the remainder term associated with the k term expansion is within the acceptable margin of error, then one need proceed no further. Unfortunately, this remainder term involves all $k+1$ order partial derivatives of the equivalent variation

[9] It is straightforward to construct proper equivalent variation functions with following property: The first k terms of the Taylor-series expansion shrink to zero, the next $n - k - 1$ terms are zero, for n arbitrarily big, and the nth term is very large and carries the preponderence of the quantitative information.

function. Moreover, to obtain error bounds, the term must be maximized (in absolute value) over the line segment connecting the initial and final points in price-income space. These are difficult calculations that can require much information about demands as functions of prices and income. Yet, absent such calculations, any finite version of the McKenzie-Pearce procedure can yield entirely misleading results.

In contrast, the theory developed in *CSWA* provides closed-form expressions that bound the errors involved in approximating the equivalent and compensating variations by means of consumer's surplus. The only data required to calculate the bounds are the consumer's income, the surplus measures themselves, and bounds on the income elasticities of demand. If the upper and lower bounds on the approximation are large, but close to one another (this would occur if the upper and lower bounds on the income elasticities were close to one another), then they would provide an accurate welfare measure unequal to consumer's surplus. If the bounds are within the acceptable margin of error, then surplus is an appropriate welfare indicator.

Perhaps most importantly, the bounds presented here and in *CSWA* have intuitive economic interpretations. They quantify and validate the notion that if the real income change caused by a price movement would have only a moderate income effect on the good's demand, then the resulting consumer's surplus is an accurate welfare measure that is commensurable with nominal income.

REFERENCES

Tom Apostol, *Mathematical Analysis*, Reading 1957.

G. McKenzie, "Consumer's Surplus Without Apology: Comment," *Amer. Econ. Rev.*, June 1979, *69*, 465–68.

______ **and I. F. Pearce,** "A New Operational Procedure for Evaluating Economic Policies," *Rev. Econ. Stud.*, Oct. 1976, *43*, 465–68.

R. Willig, (1973a) "Consumer's Surplus: A Rigorous Cookbook," tech. rept. no. 98, econ. series, IMSSS, Stanford Univ 1973.

______, (1973b) "Welfare Analysis of Policies Affecting Prices and Products," memo. no. 153, Center Res. Econ. Growth, Stanford Univ. 1973.

______, (1976a) "Consumer's Surplus Without Apology," *Amer. Econ. Rev.*, Sept. 1976, *66*, 589–97.

______, (1976b) "Integrability Implications for Locally Constant Demand Elasticities," *J. Econ. Theory*, June 1976, *12*, 391–401.

A Note on Optimal Taxation and Administrative Costs

By SHLOMO YITZHAKI*

The literature on optimal commodity taxation[1] deals with the tax rates that minimize the deadweight loss (or excess burden) of a tax system that raises a given amount of tax liabilities. A crucial assumption in this literature is that the number of taxed commodities is both given and less than the total number of commodities.[2] This assumption partitions the set of commodities into two sets, the taxed and the untaxed. Thus, by imposing a tax the government changes the prices of every taxed commodity relative to the untaxed commodities. The tax causes a substitution effect unless the demand for the taxed commodities is completely inelastic. Optimal taxation theory finds the tax rates that minimize the effect of this substitution effect. Obviously, the minimum deadweight loss is a nonincreasing function of the size of the set of taxed commodities.[3] In the extreme case where all commodities are taxed they can all be taxed at the same rate, so that relative prices do not change and we end up with a zero deadweight loss.

One major factor which prevents an increase in the number of taxed commodities is the administrative cost of taxation. By taking this into consideration, one can build a model of optimal taxation in which the number of taxable commodities is a decision variable. In such a model the target of optimal taxation is to minimize the social cost of taxation. The social cost of taxation is the sum of the administrative cost and the deadweight loss caused by the tax system. The policy instruments are the tax rates and the composition of the taxed commodities. This model will include a lump sum tax as a special case. It occurs where all commodities are taxed and therefore the optimum policy is to tax them at a flat rate.[4] The lump-sum tax is optimal only if the administrative cost associated with it is less than the social cost of any other possible tax.

In this paper I present the solution to a simple model of optimal taxation which includes the administrative cost of taxation; it is based on a Cobb-Douglas utility function. The major conclusions are: 1) Taking into account the administrative costs, the problem of optimal taxation ceases to be a problem in the theory of the second best (see Frank Hahn for a similar view). 2) The optimum level of administrative costs is an increasing function of the tax revenue, that is, the greater the government expenditure, the greater the administrative cost. 3) Define the marginal cost of administration as the additional outlay needed in order to raise an additional dollar of tax revenue. Define the marginal excess burden as the change in the excess burden due to a change in tax rates that increases the tax revenue by $1. Then in an optimum solution the marginal administrative cost and the marginal excess burden would be equal. This outcome enables us to use the marginal excess burden as a test of the optimality of the tax administration. Another empirical implication of the model is that the larger the public sector, and thus the tax revenue, the more taxes there should be.

*The Hebrew University and Falk Institute, Jerusalem. This paper was written during my stay at Harvard as a postdoctoral fellow. I am grateful to Martin Feldstein, Ariel Pakes, Eytan Sheshinski, Joel Slemrod, an anonymous referee, and the managing editor for helpful comments.

[1]For a recent review of the optimal taxation literature, see Agnar Sandmo.

[2]One way to introducing such an assumption is to assume that only transactions can be taxed (see Peter Diamond and James Mirrlees).

[3]This is always true since the set of optimum tax rates with n taxed commodities is included in the possible solutions for the case of $n + 1$ taxed commodities.

[4]The main argument against the practical use of other types of lump sum tax such as a poll tax is that in order to impose it, it must be proved that the assessee has (or had) the ability to pay. (For example, a tax cannot be greater than the assessee's resources.) A poll tax is feasible only when it is known that every assessee can afford it.

Section I presents the solution to the optimal taxation problem when there are no administrative costs. In Section II, the model of optimal taxation in the presence of administrative costs is presented and solved. Section III discusses the quantitative relationship between the administrative costs and the deadweight loss of taxation. In Section IV, I comment on the problems involved in generalizing the analysis.[5]

I. Optimum Tax Rates and the Cobb-Douglas Utility Function

Assume a representative agent with a linearly homogeneous Cobb-Douglas utility function. The problem is to find the tax rates that minimize the agent's utility loss for a given amount of tax collected, T. Following the literature on excess burden and optimal taxation (Diamond and Mirrlees), it is assumed that the prices faced by the producers P_i $(i = 1, \ldots, n)$ are constant, and that the government returns the tax revenue T to the consumer as a lump sum subsidy.

Throughout the paper the following notation is used:

$V^p = V(I, p_1, \ldots, p_n)$ is the level of the indirect utility function in the absence of taxes; I may be interpreted as the value of time plus other incomes of the agent.

$V^q = V(y, q_1, \ldots, q_n)$ is the level of the indirect utility function in the presence of indirect taxes and the income transfer to the consumer; $y = I + T$ and $q_1, \ldots, q_n$ are the prices faced by the consumer, $q_i = (1 + t_i)p_i$.

$V^y = V(y, p_1, \ldots, p_n)$ is the level of the indirect utility function where prices are producer prices and consumer income includes the transfer.

Using the indirect utility function with constant returns to scale and two taxed commodities (for simplicity), the optimum tax rates are the solution of the following problem:

$$(1) \qquad \max_{t_1, t_2} V^q = Ay \prod_{i=1}^{n} (q_i^{-\alpha_i})$$

subject to $t_1 p_1 x_1 + t_2 p_2 x_2 = T$

where $x_i = \alpha_i(y/q_i)$ is the demand f[or] commodity i and we assume $\Sigma\alpha_i = 1$. (Th[e] assumption $\Sigma\alpha_i < 1$ does not change t[he] outcome.) Using the relationship $q_i = (1 [+] t_i)p_i$ $(i = 1, 2)$ we can rewrite (1) as

$$(2) \qquad \max_{t_1, t_2} V^q = V^y(1 + t_1)^{-\alpha_1}(1 + t_2)^{-\alpha_2}$$

subject to

$$\left(\alpha_1 \frac{t_1}{1 + t_1} + \alpha_2 \frac{t_2}{1 + t_2}\right) y - T = 0$$

The solution of (2) is $t_1 = t_2$.[6] This outcome holds for any number of taxed commodities, that is, all taxed commodities should be subject to the same rate. Using this property of the solution of (2) we can write the indirect utility function as a function of two aggregate commodities: the taxed and the untaxed, i.e., $V^q = V^y(1 + t)^{-\gamma}$, where $\gamma = \Sigma_{i \in M}\alpha_i$, M is the set of taxed commodities, and t is the tax rate. We say that the size of the tax base γ has increased when the number of taxed commodities increases and V^q is a decreasing function of both γ and t.

II. Administrative Costs of Taxation

Taxation requires an administration to handle both the assessment and the collection of taxes. The size of the administration depends on such features as ability to locate the commodity, existence of a market for it, and the number of assessees. Let us assume that associated with each taxed commodity i there is a constant administrative cost C_i, which represents the need for administration and regulation to handle the tax assessment.[7]

[5]For a taxation model with an administrative cost in a broader context, see Walter P. Heller and Karl Shell.

[6]For a general discussion on the conditions which lead to an optimum tax which is uniform, see Sandmo.

[7]The existence of tax evasion and avoidance suggests that C_i is also a function of t_i when $\partial C_i/\partial t_i > 0$. Such an assumption will not change the nature of the model. In a more general case the administrative costs of handling a tax may be a function of other taxes. I comment on this point in the last section.

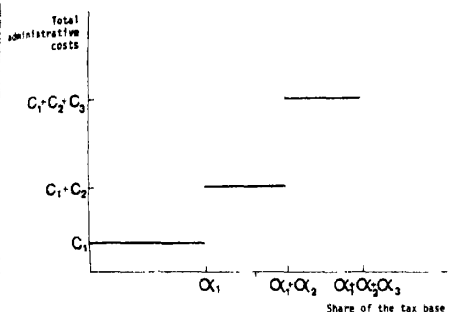



FIGURE 1. THE ADMINISTRATIVE COST FUNCTION

We list the commodities in increasing order of average administrative cost per dollar spent on each, that is, according to C_i/α_i. We can then write the function $C(\gamma) = \Sigma_{i\epsilon M} C_i$ where C is the total administrative cost associated with the tax base γ ($= \Sigma_{i\epsilon M}\alpha_i$) and M is the set of taxable commodities; C is a nondecreasing step function by construction. This is shown graphically in Figure 1.

Assume for simplicity that the number of commodities is large so that we may take a continuous approximation to $C = C(\gamma)$, $\gamma \leq 1$. The government's problem is to minimize the total utility loss caused by taxation by choosing the optimum size of the tax base γ^* and the optimum tax rate t^*, for a given amount of tax revenue T.

Formally the problem is

$$\text{(3)} \qquad \max_{t,\gamma} L = V^y(1+t)^{-\gamma}$$

subject to the government's budget constraint

$$\text{(4a)} \qquad \gamma y t(1+t)^{-1} - C(\gamma) - T = 0$$

where $y = I + T$; $V^y = V(y, p_1, \ldots, p_n)$.

The first-order condition with respect to t is

$$\text{(4b)} \qquad \frac{\partial L}{\partial t} = -\frac{\gamma}{1+t}V^q - \lambda\frac{\gamma y}{(1+t)^2} = 0$$

where λ is the Lagrange multiplier. The first part of this expression is the marginal loss of utility due to the tax rate. The second is the marginal increase in revenue multiplied by the marginal utility of tax revenue (λ).

The first-order condition with respect to γ is

$$\text{(4c)} \qquad \frac{\partial L}{\partial \gamma} = -\log(1+t)V^q + \lambda\left[\frac{t}{1+t}y - C'(\gamma)\right] = 0$$

where $C'(\gamma) = \partial C/\partial\gamma$ is the marginal administrative cost. The first part of this expression represents the marginal loss of utility due to the increase in the size of the tax base. The second is the net increase in net revenue (marginal revenue *minus* marginal administrative cost) multiplied by the marginal utility of revenue.

The optimum combination of γ and t can be derived from (4b) and (4c) as

$$\text{(5)} \qquad \frac{t - \log(1+t)}{1+t} = \frac{C'(\gamma)}{y}$$

I shall refer to (5) as the optimality condition. It states that at the margin the cost (in terms of utility loss) of raising the marginal revenue dollar by changing γ must be equal to the cost of raising it by changing t.

The budget constraint (4a) can be written as

$$\text{(6)} \qquad t = \frac{C(\gamma) + T}{\gamma y - C(\gamma) - T}$$

provided that the denominator is greater than zero, which will be so when gross tax revenue $C(\gamma) + T$ is less than total expenditures on taxed commodities.

Since $C''(\gamma) = \partial^2 C/\partial\gamma^2 > 0$, this condition bounds the feasible size of the tax base from above and from below. Denote these bounds as γ_{max} and γ_{min}, respectively. They state that there are limits on the size of the taxable sector, and that these limits are a function of the size of net tax revenue T. In particular, the lower bound states that the tax base cannot be smaller than the tax revenue that the government intends to collect. (For example, a salt tax may be feasible when T is small enough; as T increases the tax base must grow and other taxes must be added.)[8] As for the

[8]The Cobb-Douglas model precludes rejection of taxes when T is increasing. A more general utility function

upper bound γ_{max}, two cases must be distinguished: $\gamma_{max} < 1$ and $\gamma_{max} > 1$. The case $\gamma_{max} \geq 1$ implies that the upper bound is not relevant to our discussion, for it lies outside the feasible range of γ ($0 < \gamma \leq 1$). The case $\gamma_{max} < 1$ implies that it is not feasible to tax all commodities because administrative costs exceed the economy's resources. The distinction between the two cases depends on whether $C(1)$ is greater or smaller than I.

Differentiating (5) we get

$$\frac{d\gamma}{dt} = \frac{\log(1+t)}{(1+t)^2}\frac{y}{C''(\gamma)} > 0 \tag{7}$$

which states that the optimum tax rate and the optimum size of the tax base will move in the same direction. That is, if the base is made larger, the tax rate must be raised in order to satisfy the optimality condition.

Next, by differentiating the budget constraint (6) we obtain

$$\frac{d\gamma}{dt} = \frac{(\gamma y - C(\gamma) - T)^2}{(\gamma C'(\gamma) - C(\gamma) - T)y} \tag{8}$$

The assumption that $C'' > 0$ ensures that $\gamma C'(\gamma) - C(\gamma)$ is an increasing function of γ. Hence there exists γ^0 such that

$$\frac{d\gamma}{dt} = \begin{cases} < 0 & \gamma < \gamma^0 \\ > 0 & \gamma > \gamma^0 \end{cases} \tag{9}$$

Figure 2 presents the optimality condition and the budget constraint for a given T. If there exist feasible solutions of (3), there will be a unique optimum solution, since V^q is a decreasing function of both γ and t. The solution is represented by (t^*, γ^*) in the figure. The point B represents the point of minimum utility.[9] (Consequently we must check second-order conditions in order to ensure a maximum). Surprisingly, the lump sum ($\gamma = 1$) is not the optimum tax. In fact it will be the worst of all feasible taxes that satisfy the optimality condition.

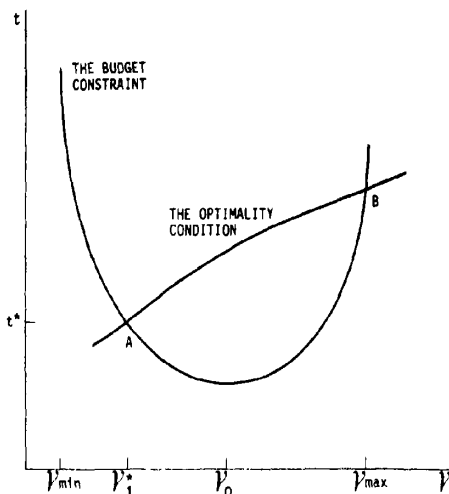



FIGURE 2 DETERMINATION OF THE OPTIMUM γ AND t

The explanation of this outcome is that the government, by changing the size of the tax base, can trade off losses due to administrative costs against excess burden. An optimal policy is one that equates the marginal loss due to administrative cost with that due to excess burden. A lump sum tax minimizes the loss due to the excess burden but ignores that due to administrative costs.

An increase in T shifts the optimality condition downward[10] and the budget constraint upward.[11] Hence an increase in the net tax revenue T increases the optimum size of the tax base γ^*. However, we cannot determine the direction of the effect of an increase in T on t^*. The increase in γ^* increases the tax base and hence tends to counteract the impact of the increase in the net tax revenue T. To illustrate the solution to the problem let $T = 0.2$, $I = 1$, and $C = 0.1(\gamma^2)$. From (4b) and (4c) we get the optimum size of tax base as 48 percent of I (the GNP plus the value of leisure), and the optimum tax rate as 60 percent. Administrative costs are 2 percent of

does not suffer from this limitation, but it would entail the problem of what is meant by "size of the tax base" in a multicommodity context. I owe this point to an anonymous referee.

[9]The point B may occur in the irrelevant range $\gamma > 1$.

[10]The change in T increases through y, the expenditure on the taxed commodities. Hence for any given γ (and marginal administrative costs) the tax rate equating the marginal excess burden with marginal administrative cost decreases.

[11]For any given γ, the tax revenue to be raised increases, hence the tax rate increases.

l. Note that by introducing administrative costs, we have taken the problem of optimal taxation out of the realm of second best solutions and turned it back into an ordinary first best optimization problem. Lump sum taxes are a possible solution of the problem, but they are not necessarily optimal because the administrative costs may exceed the utility gain from the decrease in the excess burden. In the sense that it is in fact reducible, the excess burden exists only if the tax rate or the size of the tax base are not optimal. Taking into consideration the administrative costs of taxation leads us to consider the problem of how to find the optimum size of the tax administration.

The model assumes that government (and private) expenditure on tax administration is a loss of real resources to the economy, justified only by its ability to reduce the excess burden of taxation. The excess burden is caused by changes in the relative prices of commodities. The optimum size of the tax administration can be found if the administrative cost function C, and the excess burden can be estimated.

The lack of information on the C function can be partly overcome by evaluating the sensitivity of the excess burden to the size of the taxable sector. That is, since increasing the size of the tax base (and hence administrative costs) is justified by reducing the excess burden, we can evaluate the desirability of an increase in γ by calculating the resulting change in excess burden. This is done in the next section.

III. Administrative Costs and Excess Burden

A reduction in the excess burden is the reward for an increase in administrative costs, provided T (total net revenue) is given. An estimate of the change in excess burden due to an increase in the taxable sector is an estimate of the desirability of a larger administration.

Let $\beta = T/I$ and let $E = V^q - V^p$ be the excess burden. Then $V^q = V^p(1+\beta)^\delta[1 - (\delta/\gamma)\beta]^\gamma$, where γ is the share of the taxable sector and δ is the share of the exempted sector. Using the definition of E, we find $E = V^p\{1 - (1+\beta)^\delta[1 - (\delta/\gamma)\beta]^\gamma\}$.

TABLE 1—RELATIVE GAINS IN UTILITY RESULTING FROM A MARGINAL INCREASE OF THE TAX BASE

Share of the Tax Base (γ)	Ratio of Net Tax Revenue to Total Resources (β)		
	0.1	0.2	0.3
0.5	0.022	0.095	0.238
0.8	0.008	0.030	0.065
0.9	0.006	0.022	0.049

The derivative of this expression with respect to γ is

$$\frac{-E_\gamma}{Vq} = -\log \xi - \frac{(1-\xi)}{\xi} \tag{10}$$

where $\xi = (\gamma - \delta\beta)/[\gamma(1+\beta)]$ and $E_\gamma = \partial E/\partial\gamma$. Equation (10) gives us the marginal relative gain in utility resulting from an increase in the tax base for any given γ and β.

Table 1 presents selected outcomes and shows the relative marginal gain in utility resulting from a costless increase in the tax base. This is the alternative cost (in utility terms) of marginal administrative cost. From Table 1 we can derive an estimate of the costs that would make it worthwhile to increase the size of the tax base. Assume that it accounts for 80 percent of total resources (including the value of leisure), and that tax revenue accounts for 20 percent. An additional tax should be imposed if the additional administrative costs are less than $0.03a$ (where a is the share of the taxed commodity and constant marginal utility of income is assumed).

IV. Extension and Conclusions

The model shows that if the number of taxes is allowed to vary, then the optimal taxation problem is no longer a problem in the theory of the second best. It shows that the optimum size of the taxable sector (and the administrative costs of taxation) increases with the optimum tax rate. Graphically, in the case of the Cobb-Douglas utility function, we can draw the excess burden of taxation for

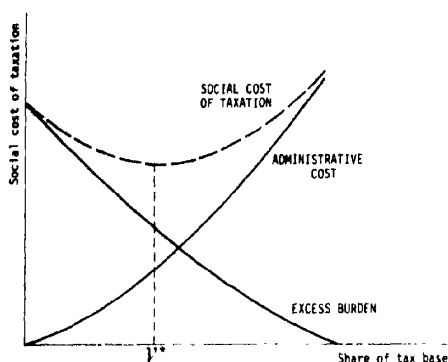



FIGURE 3. THE SOCIAL COSTS OF TAXATION FOR A GIVEN T

a given T, and the administrative cost as a function of the size of the tax base. The social cost of taxation is obtained by vertical addition of the two functions (see Figure 3). The γ that minimizes the social cost of taxation is the solution for the optimum tax problem.

The question arises as to whether these outcomes would hold in a more general model. One can think of two directions in which the model could be generalized. First, what would be the result of assuming a more general utility function $U = U(x_1, \ldots, x_n)$? The second direction is to allow the administrative cost to be a function of the set of the taxable commodities M and the tax rates. In neither of these two cases can the administrative cost and the excess burden be thought of as separate continuous functions.

Assuming a general utility function still allows us to compute, for every set of taxed commodities and given T, the excess burden of taxation. Hence we can rank the sets of taxed commodities in decreasing order of excess burden, but we cannot ensure that the administrative cost will also be in increasing order. Similarly, if we rank the set of taxable commodities according to increasing order of administrative cost, we cannot ensure that the associated excess burden would be in decreasing order. This means that in order to find the optimum set of taxed commodities we have to compute and to compare the excess burden and the administrative cost of all possible sets of taxed commodities. With n commodities, it means that we have to compare $2^n - 1$ cases.

The same type of problem arises when we try to generalize the administrative cost function. The administrative cost of taxation is a discrete function and the closure of its range is not convex. An example of potential nonconvexity is when C is not continuous in the neighborhood of the point $t_i = t_j$, i.e., when it is administratively cheaper to tax two commodities at a flat rate than at different rates. Thus there are even more than $2^n - 1$ cases to check.

Nevertheless, problems of nonconvexity and discontinuity do not affect the basic features of the simple model. The excess burden and the administrative cost can be viewed as substitutes in the general model, and the target of optimal taxation is still to minimize the social cost of taxation. This is done by choosing the tax rates and the set of taxed commodities. The necessary conditions for the optimum solution is still to equate the marginal cost of raising the tax revenue through administrative cost with the marginal cost of raising the tax revenue through the tax rates.

REFERENCES

P. A. Diamond and J. Mirrlees, "Optimal Taxation and Public Production: I—Production Efficiency," *Amer. Econ. Rev.*, Mar. 1971, *61*, 8–27; "II—Tax Rules," June 1971, *61*, 261–78.

F. H. Hahn, "On Optimal Taxation," *J. Econ. Theory*, Feb. 1973, *6*, 96–106.

W. P. Heller and K. Shell, "On Optimal Taxation with Costly Administration," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 338–45.

A. Sandmo, "Optimal Taxation and Introduction to the Literature," *J. Publ. Econ.*, July/Aug. 1976, *6*, 37–54.

Capitalization of Intrajurisdictional Differences in Local Tax Prices: Comment

By JAMES C. DYER IV AND MICHAEL D. MAHER*

In his recent article in this *Review*, Bruce Hamilton attempted to extend Peter Mieszkowski's (1969, 1972) analysis of property tax incidence. He correctly notes that Mieszkowski's model does not describe a market equilibrium. Hamilton's model accounts for the tax and benefit incidence of an intrajurisdictional property tax, and it delineates the competitive market adjustments due to capitalization effects. He then draws several conclusions regarding fundamental urban economic problems. However, Hamilton's model fails to describe a market equilibrium for the identical reason Mieszkowski's fails; potential supply adjustments generated by capitalization are neglected. The purpose of this comment is to correct the Hamilton model by including sufficient conditions to obtain a market equilibrium. Our reformulation of the model indicates that some of Hamilton's urban policy conclusions are utterly incorrect while others are valid only under very restrictive assumptions.

Consider a Hamiltonian metropolitan area comprised of three jurisdictions: 1) a homogeneous high-income housing (*HIH*) community, 2) a homogeneous low-income housing (*LIH*) community, and 3) a mixed community consisting of α percent of units of *LIH* and $(1 - \alpha)$ units of *HIH*. In the initial (pretax) equilibrium property values reflect only resource costs. A public service benefit is then provided equally to each house in every community. A proportional property tax is levied in each community based on the value of the average property in the jurisdiction. There are no net capitalization effects in the two homogeneous communities because the benefit per household equals the tax. However, the mixed community is characterized by short-run intrajurisdictional net benefit differentials (*INBD*). The *LIH* receive a fiscal surplus and the *HIH* incur fiscal burden due to average cost pricing of the public services. Hamilton concludes that the net benefit differences are capitalized into property values. Land is assumed to bear the capitalization effects because capital is mobile (p. 748, fn. 9). Hamilton defines this posttax capitalization state as efficient (in supply of housing and public service) because land value differentials exactly reflect the present value of *INBD* (see pp. 748, 750, 752).

For Hamilton's conclusion to hold we must assume the relative speeds of adjustment (i.e., mobility) of capital, renters, and land differ in that order, with capital exhibiting the greatest degree of mobility. This assumption is not inconsistent with Hamilton's model and is no doubt generally accepted.[1] It is essential to note that capitalization requires at least partial immobility of one or more factors (assuming nonzero elasticities of factor substitution).[2] Furthermore, the Hamilton "efficient" state implies land is completely immobile and thus bears the full burden of *INBD*. The immobility of land eliminates substitution possibilities so an intrajurisdictional tax imposes only income effects. An

*Texas Commerce Bank and department of economics, Rice University, respectively. We are solely responsible for the analysis but wish to thank Marian Krzyzaniak, Charles McLure, Jr., Peter Mieszkowski, and Wayne Thirsk for their criticisms.

[1]By "renters" we mean both actual renters and homeowners as consumers of housing services. The term appears to be equivalent to Hamilton's "occupant" (p. 751). The expression "mobility of land" refers to sectoral mobility—land can be used for *LIH* or *HIH*. However, land is obviously jurisdictionally immobile. Housing capital is subject to filtering and consequently sectorally mobile. New housing is both sectorally and jurisdictionally mobile.

[2]There is little harm in speaking of renters as factors if it is understood that for this paper renter mobility is synonymous with elasticity of housing service demand, and equalization of net returns is synonymous with equalization of housing rents.

efficient tax by definition does not induce substitutions. Of course any tax is efficient in the short run when behavior cannot be adjusted due to immobility.

The immediate posttax cannot represent a competitive equilibrium because as Hamilton notes land price differentials will encourage landowners to supply more *LIH*. The demand for *LIH* is stimulated by landowners offering some fraction of the potential fiscal surplus to renters. In the Hamiltonian long-run equilibrium, land values are equal (per lot), *INBD* are no longer capitalized into land values, but renters of *LIH* pay less for the public service than renters of *HIH* (see Hamilton, p. 751). However, for this to be an equilibrium, renters must be unable to bid away the public service rent differentials and thus renters must be immobile relative to capital and land. But this contradicts the relative adjustment speeds which we postulated were necessary for the posttax capitalization state. Theoretically the capitalization state and the "equilibrium" state are identical—in the former *INBD* is capitalized into land values, in the latter they are capitalized into housing rentals. Both states require one factor to be immobile. Hamilton has not characterized a *long-run equilibrium* where all factors are mobile and earning the same net of tax rate of return across jurisdictions.

If Hamilton had completed the analysis he would have determined that land values, house values, and rents do not vary according to public service costs in the long-run equilibrium. There are no *INBD* and, most importantly, all communities must be homogeneous. The rent differential of Hamilton's competitive equilibrium is an incentive for *LIH* renters to demand more *LIH* in the mixed community just as the land price differential encourages landowners to supply more *LIH*. With renters mobile, land and rental differentials will disappear only when all communities have homogeneous housing.[3]

[3]We assume that *HIH* communities are perpetually homogeneous because of zoning, or because *HIH* consumers are able to bid land rent high enough to prevent *LIH* from seeking to enter the *HIH* community to capture positive *INBD*.

Any heterogeneous community employing Hamilton's average cost pricing of the public service, no matter what the proportion of the housing mix, must have some *INBD* which is inconsistent with a long-run competitive equilibrium.

Hamilton's capitalization model and his equilibrium state led him to conclude that tax-base deterioration does not explain cumulative central city decay (p. 745), and that zoning or a head tax is necessary to maintain the efficient condition (p. 752). Our reconstruction of Hamilton's model indicates his results are valid only under certain very restrictive conditions. Consider first the decay problem. With capital relatively more mobile than renters and land immobile, land values capitalize *INBD*. Therefore as Hamilton observes, there will be no tax base deterioration and no central city-suburb migration. However, this conclusion contradicts the Hamiltonian competitive equilibrium where *LIH* has expanded in the mixed community relative to *HIH*. In addition, as the correct formulation of the long-run equilibrium indicates, there will be complete decay of the mixed community. There are no economic incentives for *HIH* to remain in the mixed jurisdiction. Ultimately, the mixed jurisdiction must become *LIH* homogeneous. Ironically Hamilton's capitalization model (given average cost pricing) guarantees cumulative central city decay.

As to the problem of maintaining his efficient state, Hamilton proposes zoning as a means of preventing any inefficient overexpansion of *LIH*. Unfortunately, the zoning theoretically precludes *any* expansion of *LIH*. A community will find it in its interest to prohibit the addition of even one *LIH* since all existing property owners will suffer capital losses from the tax base deterioration. Clearly, zoning induces substitution effects and cannot maintain the Hamilton efficient state in anything but the short run, because zoning would prohibit even efficient expansion of *LIH*. Hamilton suggests replacing the local property tax with a head tax so as to eliminate *INBD* and the need for zoning (p. 751). However, Hamilton's head tax is

efficient only in the unlikely case of equality of public service benefits across households.[4] In more realistic cases with nonzero elasticity of public service consumption with respect to property values, the head tax causes *INBD* and incentives for inefficient supply responses. For example, assume public service benefits are proportional to property value. As in the Hamilton example, *HIH* (including land) is valued at $40,000 and the *LIH* at $20,000. A *HIH* and a *LIH* receive $2,000 and $1,000 in public services, respectively. For a mixed community of half *LIH* ($\alpha = .5$) the head tax equals $1,500 per property unit. The head tax results in *INBD* but now *LIH* incurs a fiscal burden and *HIH* receives a fiscal surplus. Competitive forces will cause the inefficient overexpansion of *HIH*. Interestingly in this particular illustration, an efficient equilibrium results if Hamilton's average cost property tax method is applied. The average property value of $30,000 yields a tax rate of .05. A *LIH* pays $1,000 and a *HIH* pays $2,000 in tax which just equals their public service benefits. There are no *INBD*.

Benefit taxation is the only means of obtaining a truly efficient equilibrium. The head tax in Hamilton's paper is efficient because by the design of his model the head tax finances "head benefits," that is, it is a benefit tax. Similarly, in our proportional benefits case the average cost pricing tax is by design a benefit tax. However, in general neither the head tax nor the average cost pricing property tax will be neutral. Both taxes are redistributive. Attempts to redistribute income at the local level will generate competitive adjustments that cause these efforts to fail. Moreover, intrajurisdictional redistribution is an incentive for housing segregation. If *HIH* is subsidizing *LIH*, the *HIH* will flee the jurisdiction. Benefit taxation per se will not induce migration. Hence, local governments should, as much as possible, rely on benefit taxes and user fees as prices for services provided (see Charles McLure; Richard Musgrave and Peggy Musgrave).

Finally, Hamilton raises the issue of the relationship between capitalization and horizontal equity. In essence he argues that capitalization is a sufficient condition for horizontal equity. This proposition needs to be clarified. The imposition of a tax or tax change will result in instantaneous capitalization of any intrajurisdictional differentials in net tax benefits. The capitalization alters the current housing values in the mixed community vis-à-vis similar housing in the homogeneous communities. Because landowners in equal pretax positions incur unequal tax burdens, the capitalization does create horizontal inequities. Consequently, Hamilton's equivalence of capitalization and horizontal equity cannot refer to the tax burdens relative to pretax housing values. The horizontal equity cited by Hamilton must refer to a comparison of posttax housing values, with immobile factors. Once again the mobility assumption plays a critical role. Immobile factors guarantee full capitalization of the fiscal burdens. This equalizes the price of housing and public services. When the tax burdens are computed relative to these posttax values, the burdens are horizontally equitable. However, with mobile factors the landowners in the mixed communities can escape their fiscal burdens by moving to a homogeneous community. The *HIH* in the mixed community is worse off than if it is located in the homogeneous community. Therefore, the tax-induced capitalization creates horizontal inequity. As we have noted, the long-run equilibrium is characterized by homogeneous communities. There will be no *INBD* to be capitalized, and hence will be no horizontal inequity. Capitalization and horizontal equity are definitely not "two ways of saying the same thing" (Hamilton, p. 744).

Hamilton has extended the Mieszkowski property tax analysis by including the expenditure aspect and short-run capitalization effects. However, his supply adjustment story is incomplete and several of his implications

[4]A head tax is unlikely to be an efficient pricing mechanism for even *pure public goods* since per capita benefit allocation necessitates a zero elasticity of marginal utility of income (see Henry Aaron and Martin McGuire). Empirical estimates indicate an elasticity greater than one (see Shlomo Maital).

are erroneous. We have demonstrated 1) immobility is necessary for capitalization, 2) complete mobility leads to interjurisdictional migration and eventually to totally segregated communities, and 3) a benefit tax is the only efficient means of local taxation.

REFERENCES

H. Aaron and M. McGuire, "Public Goods and Income Distribution," *Econometrica*, Nov. 1970, *38*, 907–20.

B. W. Hamilton, "Capitalization of Intrajurisdictional Differences in Local Tax Prices," *Amer. Econ. Rev.*, Dec. 1976, *66*, 743-53.

S. Maital, "Apportionment of Public Good Benefits to Individuals," *Publ. Finance*, No. 3, 1975, *30*, 392–413.

C. E. McLure, Jr., "Revenue Sharing: Alternative to Rational Fiscal Federalism?," *Publ. Pol.*, Summer 1971, *19*, 457–78.

P. M. Mieszkowski, "Tax Incidence Theory: The Effects of Taxes on the Distribution of Income," *J. Econ. Lit.*, Dec. 1969, *7*, 1103–24.

———, "The Property Tax: An Excise Tax or a Profits Tax," *J. Publ. Econ.*, Spring 1972, *1*, 73–96.

Richard A. Musgrave and Peggy B. Musgrave, *Public Finance in Theory and Practice*, 2d ed., New York 1976.

Capitalization of Intrajurisdicitional Differences in Local Tax Prices: Reply

By BRUCE W. HAMILTON*

The first point raised by James Dyer and Michael Maher in their comment is that my earlier analysis does not consider potential supply responses to capitalization effects, and that it therefore does not describe a long-run *market* equilibrium. I agree with this point and with their description of such an equilibrium. But my paper was not concerned with a pure market equilibrium; rather I was concerned with the interaction between market and public sector supply-restriction forces. In particular, I was concerned with the efficiency and distributional consequences of various possible outcomes of the interaction between these two forces. One of the points I wished to make, but which Dyer and Maher make far better, is that a pure market equilibrium either does not exist (is unstable) or has characteristics which are obviously unrealistic. Hence my interest in the supply-restriction case.

Dyer and Maher next point out that I propose zoning as a means of restoring efficiency; for undoing the excess construction of *LIH* in a free market world. They note that there is no reason to expect zoning to restore efficiency, because even efficient zoning is not in the interest of the local zoning authorities. This point is also correct. It appears in my paper beginning on the last line of page 751. Again, my purpose was not to *advocate* zoning as a means of restoring efficiency but rather to offer the basis for examining the equity and efficiency implications of the interaction between zoning and market forces. Points 2, 3, and 4 of my conclusion are addressed to this topic.

Next, Dyer and Maher note that average cost pricing is efficient only in the case where all residents have the same demand for the local public service. This is correct. It may be that Tiebout-type migration will roughly bring this about, but I ought not to have assumed it.

Finally, Dyer and Maher are correct in noting that capitalization provides horizontal equality only after any capital gains and losses brought about by any changes in capitalization are compensated. This is why my statement on the subject includes the phrase "... in a static world" (p. 744). This phrase ought to have been repeated in the conclusion.

* Johns Hopkins University

NOTES

Report of the AEA Committee on Elections: 1979

In accordance with the bylaws on election procedures, I hereby certify the results of the recent balloting and report the actions of the Nominating Committee, the Electoral College, and the Committee on Elections.

The Nominating Committee, consisting of Walter W. Heller, Chairperson, Nancy S. Barrett, Huey J. Battle, David M. Gordon, Bert G. Hickman, Irving B. Kravis, and Ralph W. Pfouts, submitted the nominations for Vice-Presidents and members of the Executive Committee. The Electoral College, consisting of the Nominating Committee and the Executive Committee meeting together, selected the nominee for President-Elect. No petitions were received nominating additional candidates.

Moss Abramovitz was chosen as nominee for President-Elect. Nominees for Vice-Presidents were Irma Adelman, Hollis B. Chenery, Arnold C. Harberger, and Jack Hirshleifer. Nominees for members of the Executive Committee were Henry J. Aaron, Samuel Bowles, Zvi Griliches, and Daniel McFadden. The Secretary prepared biographical sketches of the candidates and distributed ballots in late summer, 1978 The Committee on Elections, consisting of Ben Bolch, Chairperson, Barbara Haskew, and C. Elton Hinshaw, *ex officio*, canvassed the ballots and filed the following results:

Number of envelopes without namers for identification	235
Number of envelopes received too late	55
Number of defective ballots	39
Number of legal ballots	4,677
	5,006

On the basis of the canvass of the votes, I certify that the following persons have been duly elected to the respective office:

President-Elect (for a term of one year)
Moses Abramovitz
Vice-Presidents (for a term of one year)
Irma Adelman
Jack Hirshleifer
Members of the Executive Committee
(for a term of three years)
Henry J. Aaron
Zvi Griliches

BEN W. BOLCH, *Chair*

Nominations for AEA Officers: 1980

The Electoral College on March 16 chose William J. Baumol as nominee for President-Elect of the American Economic Association in the balloting to be held in the autumn of 1979. Other nominees (chosen by the 1979 Nominating Committee) are: Vice-President (two to be elected), Carl F. Christ, Irving B. Kravis, H. Gregg Lewis, and Lionel McKenzie; for members of the Executive Committee (two to be elected), Jagdish N. Bhagwati, Martin S. Feldstein, Robert E. Lucas, Jr., and Walter Y. Oi.

Under a change in the bylaws as described in the *Papers and Proceedings* of this *Review*, May 1971, page 472, additional candidates may be nominated by petition, delivered to the Secretary by August 1, including signatures and addresses of not less than 6 percent of the membership of the Association for the office of President-Elect, and not less than 4 percent for each of the other offices. For the purpose of circulating petitions, address labels will be made available by the Secretary at cost.

1980 Nominating Committee of the AEA

In accordance with Section IV, paragraph 2, of the bylaws of the American Economic Association as amended in 1972, President-Elect Moses Abramovitz has appointed a Nominating Committee for 1980 consisting of Lawrence Klein, Chairman; Herman Daly, David Kendrick, Lester Thurow, Sherwin Rosen, Anna Schwartz, and Murray Wiedenbaum. Attention of members is called to the part of the bylaw reading, "In addition to appointees chosen by the President-Elect, the Committee shall include any other member of the Association nominated by petition including signature and addresses of not less than 2 percent of the members of the Association, delivered to the Secretary before December 1 No members of the Association may validly petition for more than one nominee for the Committee The names of the Committee shall be announced to the membership immediately following its appointment and the membership invited to suggest nominees for the various offices to the Committee."

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be

received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

The eighth annual conference of the Economics Department of the City College of the City University of New York, scheduled for May 15, 1980, will be concerned with the changing economic base of New York City Keynote addresses will be made by distinguished invited speakers Additional papers on any aspect of this subject will be welcomed Attendance and discussion at this conference will be open to economists, government officials, and executives of business firms. The conference papers will be published in a symposium volume. Inquiries about the conference should be addressed to the Conference Chairman, Professor Benjamin J. Klebaner, Department of Economists, City College of C.U N.Y., New York, NY 10031.

The next annual meeting of the History of Economics Society will be held April 17–19, 1980, at the Kress Library of Business and Economics, Harvard University. Suggestions for topics and speakers are invited as are proposals for papers. They may be sent to the President-Elect of the Society, William D Grampp, Department of Economics, University of Illinois, Chicago, IL 60680. It will be helpful if they are received before December 1, 1979.

The National Tax Association-Tax Institute of America announces the 1978 award winners in the annual competition for outstanding doctoral dessertations in government finance and taxation The $1,000 first prize award was won by Charles F. Revier of M.I.T. (now at Colorado State University) with his entry, "Incidence of Intermetropolitan Differentials in Property Taxes on Rental Housing." Honorable mention awards of $500 each were won by David M. de Ferranti of Princeton University (now at U.S. Department of Agriculture), "Tests of Seven Hypotheses on Welfare Dependency and Family Disintegration" and Michael Lea of the University of North Carolina (now at Cornell University), "A Simultaneous Equations Model of Residential Property Values and Local Public Expenditure Determination." The members of the 1978 Selection Committee were Professors Roy Bahl, George Break, Arthur D. Lynn, Oliver Oldman, and James A. Papke. Information on the 1979 awards competition may be obtained from Professor James A. Papke, Department of Economics, Krannert Graduate School of Management, Purdue University, West Lafayette, IN 47907.

Members of the NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics are pleased to announce the institution of an annual Leonard J. Savage Award of $500 for an outstanding doctoral dissertation in the area of Bayesian econometrics and statistics. To be considered for the 1979 Savage Award, a doctoral dissertation must be submitted by the dissertation supervisor before July 1, 1979, accompanied by a short letter from the supervisor summarizing the main results. Dissertations completed after January 1, 1976 are eligible, and should be sent to Professor Arnold Zellner, Graduate School of Business, University of Chicago, 5836 S. Greenwood Avenue, Chicago, IL 60637. An Evaluation Committee will be appointed by the board of the Leonard J Savage Memorial Trust Fund (M. H. DeGroot, S. E. Fienberg, S. Geisser, J. B. Dadane, E. E. Leamer, J W Pratt, and A. Zellner, Chairman).

Past volumes of the *American Economic Review* available, 1949 to present. Sold privately or gift to library. Contact Dean R. J. Ward, Southeastern Massachusetts University, N. Dartmouth, MA 02947.

The Association of Indian Economics Studies will hold its third conference, August 10–12, 1979, at Manhattan College, Riverdale, New York. Economists interested in giving papers should send a one-page abstract to Professor Suresh Desai, Chairperson, Program Committee, Department of Economics, Montclair State College, Upper Montclair, NJ 07043.

Brooklyn College, School of Social Science (Department of History), jointly with the Center for European Studies (East European Section), C.U.N.Y., is holding the following international multidisciplinary conferences: "War and Society in East Central Europe during the Eighteenth Century," December 3–5, 1979; "Inflation Through the Ages," May 1980; "War and Society in East Central Europe during the Ninteenth Century," December 1980. Interested scholars should contact Bela K. Kiraly, Program Director, P.O. Box 568, Highland Lakes, NJ 07422 (or telephone 201+764-4376).

The International Institute of Public Finance will hold its thirty-fifth Congress of Public Finance in Taormina, Italy, September 10–14, 1979. The topic will be "Tax Reforms (Analytical and Empirical Aspects)." Inter-

ested persons are encouraged to contract the chairman with respect to presenting papers or participating. Write to Chairman of the Scientific Committee, Professor Francesco Forte, University of Torino, Corso Francia, 7, Torino, Italy. For other scientific and institutional matters, write to the President of IIPF, Professor Horst Claus Recktenwald, Friedrich-Alexander-Universität Erlangen-Nürnberg, Lange Gasse 20, 8500 Nürnberg, Germany.

The School of Urban and Public Affairs at Carnegie-Mellon University, with the support of a training grant from the Center for Studies in Crime and Delinquency of the National Institute of Mental Health, is offering a postdoctoral program in Quantitative Methods in Criminal Justice. This program is intended to bring together specialists in criminology, sociology, political science, social psychology, criminal justice, statistics, operations research, management science, and econometrics. In addition to a stipend, all training costs and research resources are provided by the training grant. A limited number of predoctoral fellowships are also available to individuals who have already completed at least two years of graduate study Participation in the program can begin in July 1979. For further information and application forms, please write to Professor Alfred Blumstein, School of Urban and Public Affairs, Carnegie-Mellon University, Pittsburgh, PA 15213

The University of Missouri-Rolla will hold the sixth annual Conference and Exposition on Energy, October 16–18, 1979. The theme is "Energy Alternatives: An Assessment." Papers related to the many facets of technical, economic, political, and social developments in energy will be incorporated into appropriate sessions For more information write to the Conference Director, Dr. J Derald Morgan, Electrical Engineering Department, 122 Electrical Engineering Building, University of Missouri-Rolla, Rolla, MO 65401

Papers are invited for the third Conference on Major International Economic Issues to be held at the University of Southern California, June 13–14, 1981 The theme is "Multinational Enterprises in the World Economy." The impact of the multinationals in the rich and the poor countries with empirical observations, case studies, and policy orientation will be the main emphasis. Papers that are accepted will be published Submit proposals to Professor Nake M. Kamrany, Department of Economics, University of Southern California, Los Angeles, CA 90007.

The *Intermountain Economic Review*, a student-edited journal published in cooperation with the Department of Economics, University of Utah, was renamed the *Economic Forum*. This change is effective with the Spring 1979 issue (Vol. 10, No. 1). The basic format and emphasis of the journal will remain the same.

Social Science Quarterly announces a topical issue on Metropolitan Metamorphosis and Regional Change: The Redistribution of People and Functions in the United States. Recent trends in the movement of people, social, political, and economic activities are challenging traditional concepts of city and suburb, and raise new issues in the study of urbanization, regionalism, and related phenomena. Papers are invited from researchers interested in investigating the causes and consequences of these trends. A special issue of *SSQ* will be coedited by sociologist David Sly, political scientist Thomas Dye, economist William Serow, and geographer Wilbur Zelinsky. Papers will be accepted for review through November 1979, and should be sent to Charles M. Bonjean, Editor, *SSQ*, University of Texas, Austin, TX 78712.

There will be a session at the 1979 Economic History Association meetings in Wilmington, Delaware, September 13–15, devoted to reports on dissertation research by six to eight students who will recieve the Ph.D in economic history by the end of the summer, 1979 Individuals wishing to be considered for participation should send two copies of a 2,000-word abstract of their dissertation by June 15, 1979, the names of those selected will be announced July 1 Write to Professor Peter H Lindert, Professor Alan L Olmstead; Department of Economics, University of California-Davis, Davis, CA 95616

Call for Papers: The Rocky Mountain Conference on British Studies will hold its annual meeting, October 26–27, 1979, at the University of Colorado, Colorado Springs Paper proposals should be submitted by July 1 to Richard Cosgrove, Department of History, University of Arizona, Tucson, AZ 75721 Further information about local arrangements is available from Richard Wunderli, Department of History, University of Colorado, Colorado Springs, CO 80907.

The Institute for Socioeconomic Studies, White Plains, NY, announces the 1978 recipient of the Institute's National Service Award· Alfred J Tella, former research professor of economics at Georgetown University, and currently Special Adviser, Office of the Director, Bureau of the Census. The $5,000 award was given in recognition of Mr. Tella's contributions to economic research in the areas of labor force participation and income maintenance.

New Journal: The International Journal of Managerial Economics, May 1979. Papers and other editorial material should be sent to Dr. W. Duncan Reekie, University of Edinburgh, or to Dr. I. H. McNicholl, University of Strathclyde, Glasgow. Subscriptions and other enquiries should be forwarded to M.C.B. Publications Ltd., Keighley Road, Bradford, U.K.

The Kentucky Economic Association is seeking a limited number of contributed papers for presentation as part of the program at its annual conference in Louisville, Kentucky on October 26, 1979. Preference will be given to papers that have direct implication to the Kentucky economy. Interested persons should submit relevant papers by July 1, 1979 to Dr. James R. McCabe, School of Business, University of Louisville, Louisville, KY 40208.

The American Economic Association Committee on U S -Soviet Exchanges announces the fifth U.S.-Soviet Economic Symposium, "Long-Term Structural Change in National Economies," to be held June 10–13, 1979, at Seven Springs Farm, Mount Kisco, New York.

It is expected that the sixth symposium will be held in the *USSR*. The Committee will welcome suggestions for subjects to be discussed

Several directors of research institutes in the *USSR* have expressed willingness to participate in programs of parallel research with American economists, that is, programs in which some American economists would work on subjects similar to those being studied by Soviet economists, with the two groups meeting occasionally to compare methods and results Anyone wishing to propose a program of this sort should communicate with a member of the AEA Committee. Abram Bergson, Harvard University, John Meyer, Harvard University; Chairman Lloyd G Reynolds, Yale University.

Deaths

Donald R. Kaldor, professor of economics, Iowa State University, Oct 9, 1978

William S. Neiswanger, professor emeritus of economics, University of Illinois, Oct. 24, 1978.

Retirement

Dwight Flanders, professor emeritus of economics, University of Illinois, Aug. 31, 1977.

Visiting Foreign Scholars

Robert E Ankli, University of Guelph: visiting associate professor, department of economics, University of California-Davis, July 1, 1978.

Xavier Freixas, l'Université des Sciences Sociales, France: visiting assistant professor, department of economics, University of Illinois, Fall 1978.

Toshihiko Hayashi, Kobe University, Japan: visiting associate professor, department of economics, University of California-Davis, July 1, 1978.

A. J. H. Latham, University College of Swansea, England: visiting associate professor, department of economics, University of Illinois, Spring 1979.

Alan S. Milward, University of Manchester, England: visiting professor, department of economics, University of Illinois, Fall 1978.

Kimio Morimune, Kyoto University, Japan: visiting assistant professor, department of economics, University of Illinois, 1979.

Promotions

Gerald E. Auten: associate professor of economics, Bowling Green State University.

Paul Beckerman: assistant professor of economics, University of Illinois, Dec. 1978.

Francine Blau: associate professor of economics, University of Illinois, Aug. 1978.

Harry Brandt: senior vice président and director of research, Federal Reserve Bank of Atlanta, Jan. 1, 1979.

Charles R Chittle. professor of economics, Bowling Green State University.

Edward T. Dowling· associate professor, economics department, Fordham University, Sept. 1978.

Robert F Engle. professor of economics, University of California-San Diego, July 1, 1977.

Peter Fousek. senior vice president and director of research, Federal Reserve Bank of New York, Jan. 1, 1979.

Paul F. Haas: professor of economics, Bowling Green State University.

Walter P. Heller: professor of economics, University of California-San Diego, July 1, 1978.

J. Vernon Henderson. associate professor, department of economics, Brown University, July 1, 1979.

Wallace E. Hendricks: associate professor of economics, University of Illinois, Aug. 1978.

Larry D Neal. professor of economics, University of Illinois, Aug. 1978.

Alan Olmstead· professor of economics, University of California-Davis, July 1, 1978

Scott E. Pardee: senior vice president, Federal Reserve Bank of New York, Jan 1, 1979.

Laura R. Randall: professor of economics, Hunter College, Jan. 1, 1979

J. David Reed: professor of economics, Bowling Green State University.

Edward Rice: assistant professor of economics, University of Illinois, Jan. 1979.

Brian Rungeling: professor of economics, University of Mississippi, July 1978.

Todd Sandler: professor of economics, University of Wyoming, July 1, 1979.

Steven A. Seelig: associate professor, economics department, Fordham University, Sept. 1978.

Rudolf Thunberg: vice president and assistant director of research, Federal Reserve Bank of New York, Jan. 1, 1979.

Thomas S. Ulen: assistant professor of economics, University of Illinois, Jan. 1979.

Administrative Appointments

Arthur Kraft: associate dean, College of Business Administration, University of Nebraska-Lincoln, Dec. 1978.

Neil B. Murphy, University of Maine: senior vice president, Payment Systems, Inc., Atlanta.

David W. Penn: general manager, Wisconsin Public Power Inc., Apr. 1979.

Brian Rungeling: chairman, department of economics and finance, University of Mississippi, July 1978.

James M. Suarez: chairman, department of economics, Hunter College, Dec. 6, 1978.

Appointments

Paul B. Bennett economists, Business Conditions Division, Federal Reserve Bank of New York, Nov. 1, 1978.

Mark Browning: assistant professor of economics, University of Illinois, Jan. 1979.

Bruce F. Davie: chief tax economist, Committee on Ways and Means, U.S. House of Representative, June 1978.

Robert Feldman: economist, Developing Economies Division, Federal Reserve Bank of New York, Feb. 7, 1979.

Kenneth D. Garbade: special assistant, domestic research department, Federal Reserve Bank of New York, Dec. 29, 1978.

Roger G. Ginder: assistant professor, department of ecomonics, Iowa State University, Nov 20, 1978.

Robert E. Hall, Massachusetts Institute of Technology: senior fellow, Hoover Institution, Sept 1, 1978

Stephen H. Long assistant professor, department of economics, Syracuse University, Sept. 1978.

Steve Matthews. assistant professor of economics, University of Illinois, Aug. 1978.

Francoise Shoumaker: assistant professor of economics, University of Illinois, Aug. 1978.

Wayne M. Simon: economist, Industrial Economies Division, Federal Reserve Bank of New York, Dec. 27, 1978.

Ragnhild Sohlberg: adjunct professor, Defense Resources Management Education Center, Naval Postgraduate School, Monterey, California, Aug. 29, 1978.

Robert C. Vogel, Southern Illinois University: professor, department of economics, Syracuse University, Sept. 1978.

Leaves for Special Appointment

Martin Bronfenbrenner, Duke University: visiting scholar, Federal Reserve Bank of San Francisco, Jan.–Aug. 1979.

Warren L. Coats, Jr., International Monetary Fund visiting economist, Board of Governors of the Federal Reserve System, Dec. 1978–Nov. 1979.

James R. Millar, University of Illinois: Soviet Academy of Sciences, Moscow, Spring 1979.

Leonard J. Mirman, University of Illinois: CORE, Belgium, Nov. 1978–June 1979.

R. Robert Russell, University of California-San Diego deputy director, Council of Wage and Price Stability, 1978–79.

John Scadding, University of Toronto consultant, Federal Reserve Bank of San Francisco, Dec. 1978–Sept 1979.

Robert Schoeplein, University of Illinois: Social Security Administration, 1978–79.

Takashi Takayama, University of Illinois: Oak Ridge National Laboratory, 1978–79.

Paul J. Uselding, University of Illinois: Guggenheim fellowship, Harvard University, 1978–79

Resignations

John W. Ball, Iowa State University: Etna Insurance Company, Nov. 30, 1978.

Craig O'Riley, Iowa State University. Iowa Department of Transportation, Nov. 17, 1978

Ronald Raikes, Iowa State University, Feb. 28, 1979

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1979 issue of the *Review* will carry the seventy-sixth list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1979. This announcement is an invitation to send us information for the preparation of the list. This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1979 issue the names of those who will have been awarded the doctoral degree since June 1978 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3 x 5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ______

Name: LAST NAME IN CAPS: First Name, Initial

Institution Granting Degree: ______

Degree Conferred (Ph.D. or D.B.A.) ______ Year ______

Dissertation Title: ______

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories:

1--Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment: her new title (if any), and the date at which the change will occur.

C. Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

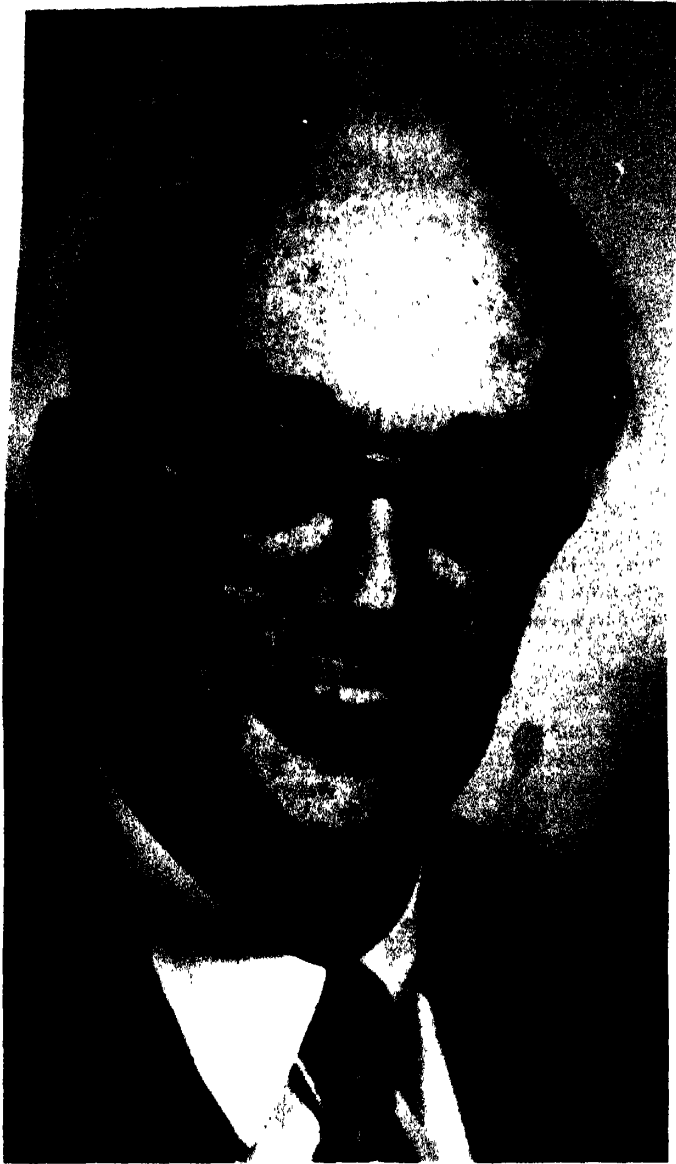
This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University Providence, Rhode Island 02912

WILLIAM S. VICKREY

DISTINGUISHED FELLOW

1978

Many of us have had the experience of thinking we were the first to show the neutrality of a particular tax scheme, to prove the incentive characteristics of a particular bidding institution, to deduce the redistributive implications of the expected utility hypothesis, to invent a demand revealing process, and so on, only to find that William S. Vickrey had done it earlier—sometimes much earlier—and whereas our "original contribution" may have contained a minor or even a substantive error, Vickrey had done it correctly. Some great scholars receive recognition from the beginning, but, inscrutably, with others it takes a little longer. His numerous works, appearing in all the leading journals in economics, law, operations research, finance, and taxation, contain many seminal contributions, and many more that would have been seminal but for the fact that the profession was not yet ready for his ideas. Thus, his "Counterspeculation, Auctions and Competitive Sealed Tenders," *Journal of Finance*, 1961, (1) invents a class of demand revealing processes for private goods, (2) develops with clarity the important concept of incentive compatibility, and (3) operationalizes these theoretical insights in the form of realizable auctioning institutions. Only later after the profession had discovered solutions to the "free-rider" problem was it possible to adequately appreciate Vickrey's astonishing precursory insights.

We are proud to recognize the creative, inspirational, and persistently operational character of William S. Vickrey's contributions to economic theory and economic policy.

William Vickrey

Rational Decision Making in Business Organizations

By HERBERT A. SIMON*

In the opening words of his *Principles*, Alfred Marshall proclaimed economics to be a psychological science:

> Political Economy or Economics is a study of mankind in the ordinary business of life; it examines that part of individual and social action which is most closely connected with the attainment and with the use of the material requisites of wellbeing.
>
> Thus it is on the one side a study of wealth; and on the other, and more important side, a part of the study of man. For man's character has been moulded by his every-day work, and the material resources which he thereby procures, more than by any other influence unless it be that of his religious ideals.

In its actual development, however, economic science has focused on just one aspect of man's character, his reason, and particularly on the application of that reason to problems of allocation in the face of scarcity. Still, modern definitions of the economic sciences, whether phrased in terms of allocating scarce resources or in terms of rational decision making, mark out a vast domain for conquest and settlement. In recent years there has been considerable exploration by economists even of parts of this domain that were thought traditionally to belong to the disciplines of political science, sociology, and psychology.

I. Decision Theory as Economic Science

The density of settlement of economists over the whole empire of economic science is very uneven, with a few areas of modest size holding the bulk of the population. The economic Heartland is the normative study of the international and national economies and their markets, with its triple main concerns of full employment of resources, the efficient allocation of resources, and equity in distribution of the economic product. Instead of the ambiguous and over-general term "economics," I will use "political economy" to designate this Heartland, and "economic sciences" to denote the whole empire, including its most remote colonies. Our principal concern in this paper will be with the important colonial territory known as decision theory. I will have something to say about its normative and descriptive aspects, and particularly about its applications to the theory of the firm. It is through the latter topic that the discussion will be linked back to the Heartland of political economy.

Underpinning the corpus of policy-oriented normative economics, there is, of course, an impressive body of descriptive or "positive" theory which rivals in its mathematical beauty and elegance some of the finest theories in the physical sciences. As examples I need only remind you of Walrasian general equilibrium theories and their modern descendants in the works of Henry Schultz, Samuelson, Hicks, and others; or the subtle and impressive body of theory created by Arrow, Hurwicz, Debreu, Malinvaud, and their colleagues showing the equivalence, under certain conditions, of competitive equilibrium with Pareto optimality.

The relevance of some of the more refined parts of this work to the real world can be, and has been, questioned. Perhaps some of these intellectual mountains have been

*Carnegie-Mellon University. This article is the lecture Herbert Simon delivered in Stockholm, Sweden, December 8, 1978, when he received the Nobel Prize in Economic Science.

The author is indebted to Albert Ando, Otto A. Davis, and Benjamin M. Friedman for valuable comments on an earlier draft of this paper.

climbed simply because they were there—because of the sheer challenge and joy of scaling them. That is as it should be in any human scientific or artistic effort. But regardless of the motives of the climbers, regardless of real world veridicality, there is no question but that positive political economy has been strongly shaped by the demands of economic policy for advice on basic public issues.

This too is as it should be. It is a vulgar fallacy to suppose that scientific inquiry cannot be fundamental if it threatens to become useful, or if it arises in response to problems posed by the everyday world. The real world, in fact, is perhaps the most fertile of all sources of good research questions calling for basic scientific inquiry.

A. *Decision Theory in the Service of Political Economy*

There is, however, a converse fallacy that deserves equal condemnation: the fallacy of supposing that fundamental inquiry is worth pursuing only if its relevance to questions of policy is immediate and obvious. In the contemporary world, this fallacy is perhaps not widely accepted, at least as far as the natural sciences are concerned. We have now lived through three centuries or more of vigorous and highly successful inquiry into the laws of nature. Much of that inquiry has been driven by the simple urge to understand, to find the beauty of order hidden in complexity. Time and again, we have found the "idle" truths arrived at through the process of inquiry to be of the greatest moment for practical human affairs. I need not take time here to argue the point. Scientists know it, engineers and physicians know it, congressmen and members of parliaments know it, the man on the street knows it.

But I am not sure that this truth is as widely known in economics as it ought to be. I cannot otherwise explain the rather weak and backward development of the descriptive theory of decision making including the theory of the firm, the sparse and scattered settlement of its terrain, and the fact that many, if not most, of its investigators are drawn from outside economics—from sociology, from psychology, and from political science. Respected and distinguished figures in economics—Edward Mason, Fritz Machlup, and Milton Friedman, for example—have placed it outside the Pale (more accurately, have placed economics outside *its* Pale), and have offered it full autonomy provided that it did not claim close kinship with genuine economic inquiry.

Thus, Mason, commenting on Papandreou's 1952 survey of research on the behavioral theory of the firm, mused aloud:

> ... has the contribution of this literature to economic analysis really been a large one? ... The writer of this critique must confess a lack of confidence in the marked superiority, *for purposes of economic analysis*, of this newer concept of the firm, over the older conception of the entrepreneur. [pp. 221–22]

And, in a similar vein, Friedman sums up his celebrated polemic against realism in theory:

> Complete "realism" is clearly unattainable, and the question whether a theory is realistic "enough" can be settled only by seeing whether it yields predictions that are good enough *for the purpose in hand* or that are better than predictions from alternative theories.
> [p. 41, emphasis added]

The "purpose in hand" that is implicit in both of these quotations is providing decision-theoretic foundations for positive, and then for normative, political economy. In the views of Mason and Friedman, fundamental inquiry into rational human behavior in the context of business organizations is simply not (by definition) economics—that is to say, political economy—unless it contributes in a major way to that purpose. This is sometimes even interpreted to mean that economic theories of decision making are not falsified in any interesting or relevant sense when their empirical predictions of *microphenomena* are found to be grossly incompatible with the observed data. Such theories, we are told, are still realistic "enough" provided that they do not contradict aggregate observations of concern

to political economy. Thus economists who are zealous in insisting that economic actors maximize turn around and become satisficers when the evaluation of their own theories is concerned. They believe that businessmen maximize, but they know that economic theorists satisfice.

The application of the principle of satisficing to theories is sometimes defended as an application of Occam's Razor: accept the simplest theory that works.[1] But Occam's Razor has a double edge. Succinctness of statement is not the only measure of a theory's simplicity. Occam understood his rule as recommending theories that make no more assumptions than necessary to account for the phenomena (*Essentia non sunt multiplicanda praeter necessitatem*). A theory of profit or utility maximization can be stated more briefly than a satisficing theory of the sort I shall discuss later. But the former makes much stronger assumptions than the latter about the human cognitive system. Hence, in the case before us, the two edges of the razor cut in opposite directions.

In whichever way we interpret Occam's principle, parsimony can be only a secondary consideration in choosing between theories, unless those theories make identical predictions. Hence, we must come back to a consideration of the phenomena that positive decision theory is supposed to handle. These may include both phenomena at the microscopic level of the decision-making agents, or aggregative phenomena of concern to political economy.

[1]The phrase "that works" refutes, out of hand, Friedman's celebrated paean of praise for lack of realism in assumptions. Consider his example of falling bodies (pp. 16–19). His valid point is that it is advantageous to use the simple law, ignoring air resistance, when it gives a "good enough" approximation. But of course the conditions under which it gives a good approximation are not at all the conditions under which it is unrealistic or a "wildly inaccurate descriptive representation of reality." We can use it to predict the path of a body falling in a vacuum, but not the path of one falling through the Earth's atmosphere. I cannot in this brief space mention, much less discuss, all of the numerous logical fallacies that can be found in Friedman's 40-page essay. For additional criticism, see Simon (1963) and Samuelson (1963).

B. *Decision Theory Pursued for its Intrinsic Interest*

Of course the definition of the word "economics" is not important. Like Humpty Dumpty, we can make words mean anything we want them to mean. But the professional training and range of concern of economists does have importance. Acceptance of the narrow view that economics is concerned only with the aggregative phenomena of political economy defines away a whole rich domain of rational human behavior as inappropriate for economic research.

I do not wish to appear to be admitting that the behavioral theory of the firm *has been* irrelevant to the construction of political economy. I will have more to say about its relevance in a moment. My present argument is counterfactual in form: *even if* there were no present evidence of such relevance, human behavior in business firms constitutes a highly interesting body of empirical phenomena that calls out for explanation as do all bodies of phenomena. And if we may extrapolate from the history of the other sciences, there is every reason to expect that as explanations emerge, relevance for important areas of practical application will not be long delayed.

It has sometimes been implied (Friedman, p. 14) that the correctness of the assumptions of rational behavior underlying the classical theory of the firm is not merely irrelevant, but is not even empirically testable in any direct way, the only valid test being whether these assumptions lead to tolerably correct predictions at the macroscopic level. That would be true, of course, if we had no microscopes, so that the micro-level behavior was not directly observable. But we do have microscopes. There are many techniques for observing decision-making behavior, even at second-by-second intervals if that is wanted. In testing our economic theories, we do not have to depend on the rough aggregate time-series that are the main grist for the econometric mill, or even upon company financial statements.

The classical theories of economic decision making and of the business firm make very specific testable predictions about the con-

crete behavior of decision-making agents. Behavioral theories make quite different predictions. Since these predictions can be tested directly by observation, either theory (or both) may be falsified as readily when such predictions fail as when predictions about aggregate phenomena are in error.

C. *Aggregative Tests of Decision Theory: Marginalism*

If some economists have erroneously supposed that micro-economic theory can only be tested by its predictions of aggregate phenomena, we should avoid the converse error of supposing that aggregate phenomena are irrelevant to testing decision theory. In particular, are there important, *empirically verified*, aggregate predictions that follow from the theory of perfect rationality but that do not follow from behavioral theories of rationality?

The classical theory of omniscient rationality is strikingly simple and beautiful. Moreover, it allows us to predict (correctly or not) human behavior without stirring out of our armchairs to observe what such behavior is like. All the predictive power comes from characterizing the shape of the environment in which the behavior takes place. The environment, combined with the assumptions of perfect rationality, fully determines the behavior. Behavioral theories of rational choice—theories of bounded rationality—do not have this kind of simplicity. But, by way of compensation, their assumptions about human capabilities are far weaker than those of the classical theory. Thus, they make modest and realistic demands on the knowledge and computational abilities of the human agents, but they also fail to predict that those agents will equate costs and returns at the margin.

D. *Have the Marginalist Predictions Been Tested?*

A number of empirical phenomena have been cited as providing more or less conclusive support for the classical theory of the firm as against its behavioral competitors (see Dale Jorgensen and Calvin Siebert). But there are no direct observations that individuals or firms do actually equate marginal costs and revenues. The empirically verified consequences of the classical theory are almost always weaker than this. Let us look at four of the most important of them: the fact that demand curves generally have negative slopes; the fact that fitted Cobb-Douglas functions are approximately homogeneous of the first degree; the fact of decreasing returns to scale; and the fact that executive salaries vary with the logarithm of company size. Are these indeed facts? And does the evidence support a maximizing theory against a satisficing theory?

Negatively Sloping Demand Curves. Evidence that consumers actually distribute their purchases in such a way as to maximize their utilities, and hence to equate marginal utilities, is nonexistent. What the empirical data do confirm is that demand curves generally have negative slopes. (Even this "obvious" fact is tricky to verify, as Henry Schultz showed long years ago.) But negatively sloping demand curves could result from a wide range of behaviors satisfying the assumptions of bounded rationality rather than those of utility maximization. Gary Becker, who can scarcely be regarded as a hostile witness for the classical theory, states the case very well:

> Economists have long been aware that some changes in the feasible or opportunity sets of households would lead to the same response *regardless of the decision rule used*. For example, a decrease in real income necessarily decreases the amount spent on at least one commodity... It has seldom been realized, however, that the change in opportunities resulting from a change in relative prices also tends to produce a systematic response, regardless of the decision rule. In particular, the fundamental theorem of traditional theory—that demand curves are negatively inclined—largely results from the change in opportunities alone and is largely independent of the decision rule. [p. 4]

Later, Becker is even more explicit, saying, "Not only utility maximization but also many other decision rules, incorporating a wide

variety of irrational behavior, lead to negatively inclined demand curves because of the effect of a change in prices on opportunities" (p. 5).[2]

First-Degree Homogeneity of Production Functions. Another example of an observed phenomenon for which the classical assumptions provide sufficient, but not necessary, conditions is the equality between labor's share of product and the exponent of the labor factor in fitted Cobb-Douglas production functions (see Simon and Ferdinand Levy). Fitted Cobb-Douglas functions are homogeneous, generally of degree close to unity and with a labor exponent of about the right magnitude. These findings, however, cannot be taken as strong evidence for the classical theory, for the identical results can readily be produced by mistakenly fitting a Cobb-Douglas function to data that were in fact generated by a linear accounting identity (value of goods equals labor cost plus capital cost), (see E. H. Phelps-Brown). The same comment applies to the SMAC production function (see Richard Cyert and Simon). Hence, the empirical findings do not allow us to draw any particular conclusions about the relative plausibility of classical and behavioral theories, both of which are equally compatible with the data.

The Long-Run Cost Curve. Somewhat different is the case of the firm's long-run cost curve, which classical theory requires to be U shaped if competitive equilibrium is to be stable. Theories of bounded rationality do not predict this—fortunately, for the observed data make it exceedingly doubtful that the cost curves are in fact generally U shaped. The evidence for many industries shows costs at the high-scale ends of the curves to be essentially constant or even declining (see Alan Walters). This finding is compatible with stochastic models of business firm growth and size (see Y. Ijiri and Simon), but not with the static equilibrium model of classical theory.

Executive Salaries. Average salaries of top corporate executives grow with the logarithm of corporate size (see David Roberts). This finding has been derived from the assumptions of the classical theory of profit maximization only with the help of very particular *ad hoc* assumptions about the distribution of managerial ability (see Robert Lucas, 1978). The observed relation is implied by a simple behavioral theory that assumes only that there is a single, culturally determined, parameter which fixes the average ratio of the salaries of managers to the salaries of their immediate subordinates (see Simon, 1957). In the case of the executive salary data, the behavioral model that explains the observations is substantially more parsimonious (in terms of assumptions about exogenous variables) than the classical model that explains the same observations.

Summary: Phenomena that Fail to Discriminate. It would take a much more extensive review than is provided here to establish the point conclusively, but I believe it is the case that specific phenomena requiring a theory of utility or profit maximization for their explanation rather than a theory of bounded rationality simply have not been observed in aggregate data. In fact, as my last two examples indicate, it is the classical rather than the behavioral form of the theory that faces real difficulties in handling some of the empirical observations.

Failures of Classical Theory. It may well be that classical theory can be patched up sufficiently to handle a wide range of situations where uncertainty and outguessing phenomena do not play a central role—that is, to handle the behavior of economies that are relatively stable and not too distant from a competitive equilibrium. However, a strong positive case for replacing the classical theory by a model of bounded rationality begins to emerge when we examine situations involving decision making under uncertainty and imperfect competition. These situations the classical theory was never designed to handle, and has never handled satisfactorily. Statistical decision theory employing the idea of subjective expected utility, on the one hand, and game theory, on the other, have contributed enormous conceptual clarification to these kinds of situations without providing

[2]In a footnote, Becker indicates that he denotes as irrational "[A]ny deviation from utility maximization." Thus, what I have called "bounded rationality" is "irrationality" in Becker's terminology.

satisfactory descriptions of actual human behavior, or even, for most cases, normative theories that are actually usable in the face of the limited computational powers of men and computers.

I shall have more to say later about the positive case for a descriptive theory of bounded rationality, but I would like to turn first to another territory within economic science that has gained rapidly in population since World War II, the domain of normative decision theory.

E. *Normative Decision Theory*

Decision theory can be pursued not only for the purposes of building foundations for political economy, or of understanding and explaining phenomena that are in themselves intrinsically interesting, but also for the purpose of offering direct advice to business and governmental decision makers. For reasons not clear to me, this territory was very sparsely settled prior to World War II. Such inhabitants as it had were mainly industrial engineers, students of public administration, and specialists in business functions, none of whom especially identified themselves with the economic sciences. Prominent pioneers included the mathematician, Charles Babbage, inventor of the digital computer, the engineer, Frederick Taylor, and the administrator, Henri Fayol.

During World War II, this territory, almost abandoned, was rediscovered by scientists, mathematicians, and statisticians concerned with military management and logistics, and was renamed "operations research" or "operations analysis." So remote were the operations researchers from the social science community that economists wishing to enter the territory had to establish their own colony, which they called "management science." The two professional organizations thus engendered still retain their separate identities, though they are now amicably federated in a number of common endeavors.

Optimization techniques were transported into management science from economics, and new optimization techniques, notably linear programming, were invented and developed, the names of Dantzig, Kantorovich, and Koopmans being prominent in the early development of that tool.

Now the salient characteristic of the decision tools employed in management science is that they have to be capable of actually making or recommending decisions, taking as their inputs the kinds of empirical data that are available in the real world, and performing only such computations as can reasonably be performed by existing desk calculators or, a little later, electronic computers. For these domains, idealized models of optimizing entrepreneurs, equipped with complete certainty about the world—or, at worst, having full probability distributions for uncertain events—are of little use. Models have to be fashioned with an eye to practical computability, no matter how severe the approximations and simplifications that are thereby imposed on them.

Model construction under these stringent conditions has taken two directions. The first is to retain optimization, but to simplify sufficiently so that the optimum (in the simplified world!) is computable. The second is to construct satisficing models that provide good enough decisions with reasonable costs of computation. By giving up optimization, a richer set of properties of the real world can be retained in the models. Stated otherwise, decision makers can satisfice either by finding optimum solutions for a simplified world, or by finding satisfactory solutions for a more realistic world. Neither approach, in general, dominates the other, and both have continued to co-exist in the world of management science.

Thus, the body of theory that has developed in management science shares with the body of theory in descriptive decision theory a central concern with the *ways* in which decisions are made, and not just with the decision outcomes. As I have suggested elsewhere (1978b), these are theories of *how* to decide rather than theories of *what* to decide.

Let me cite one example, from work in which I participated, of how model building in normative economics is shaped by computational considerations (see Charles Holt, Franco Modigliani, John Muth, and Simon).

In face of uncertain and fluctuating production demands, a company can smooth and stabilize its production and employment levels at the cost of holding buffer inventories. What kind of decision rule will secure a reasonable balance of costs? Formally, we are faced with a dynamic programming problem, and these generally pose formidable and often intolerable computational burdens for their solution.

One way out of this difficulty is to seek a special case of the problem that will be computationally tractable. If we assume the cost functions facing the company all to be quadratic in form, the optimal decision rule will then be a linear function of the decision variables, which can readily be computed in terms of the cost parameters. Equally important, under uncertainty about future sales, only the expected values, and not the higher moments, of the probability distributions enter into the decision rule (Simon, 1956b). Hence the assumption of quadratic costs reduces the original problem to one that is readily solved. Of course the solution, though it provides optimal decisions for the simplified world of our assumptions, provides, at best, satisfactory solutions for the real-world decision problem that the quadratic function approximates. In-principle, unattainable optimization is sacrificed for in-practice, attainable satisfaction.

If human decision makers are as rational as their limited computational capabilities and their incomplete information permit them to be, then there will be a close relation between normative and descriptive decision theory. Both areas of inquiry are concerned primarily with procedural rather than substantive rationality (Simon, 1978a). As new mathematical tools for computing optimal and satisfactory decisions are discovered, and as computers become more and more powerful, the recommendations of normative decision theory will change. But as the new recommendations are diffused, the actual, observed, practice of decision making in business firms will change also. And these changes may have macro-economic consequences. For example, there is some agreement that average inventory holdings of American firms have been reduced significantly by the introduction of formal procedures for calculating reorder points and quantities.

II. Characterizing Bounded Rationality

The principal forerunner of a behavioral theory of the firm is the tradition usually called Institutionalism. It is not clear that all of the writings, European and American, usually lumped under this rubric have much in common, or that their authors would agree with each other's views. At best, they share a conviction that economic theory must be reformulated to take account of the social and legal structures amidst which market transactions are carried out. Today, we even find a vigorous development within economics that seeks to achieve institutionalist goals within the context of neoclassical price theory. I will have more to say about that a little later.

The name of John R. Commons is prominent—perhaps the most prominent—among American Institutionalists. Commons' difficult writings (for example, *Institutional Economics*) borrow their language heavily from the law, and seek to use the *transaction* as their basic unit of behavior. I will not undertake to review Commons' ideas here, but simply remark that they provided me with many insights in my initial studies of organizational decision making (see my *Administrative Behavior*, p. 136).

Commons also had a substantial influence on the thinking of Chester I. Barnard, an intellectually curious business executive who distilled from his experience as president of the New Jersey Bell Telephone Company, and as executive of other business, governmental, and nonprofit organizations, a profound book on decision making titled *The Functions of the Executive*. Barnard proposed original theories, which have stood up well under empirical scrutiny, of the nature of the authority mechanism in organizations, and of the motivational bases for employee acceptance of organizational goals (the so-called "inducements-contributions" theory); and he provided a realistic description of organizational decision making, which he characterized as "opportunistic." The numer-

ous references to Barnard's work in *Administrative Behavior* attest, though inadequately, to the impact he had on my own thinking about organizations.

A. *In Search of a Descriptive Theory*

In 1934–35, in the course of a field study of the administration of public recreational facilities in Milwaukee, which were managed jointly by the school board and the city public works department, I encountered a puzzling phenomenon. Although the heads of the two agencies appeared to agree as to the objectives of the recreation program, and did not appear to be competing for empire, there was continual disagreement and tension between them with respect to the allocation of funds between physical maintenance, on the one hand, and play supervision on the other. Why did they not, as my economics books suggested, simply balance off the marginal return of the one activity against that of the other?

Further exploration made it apparent that they didn't equate expenditures at the margin because, intellectually, they couldn't. There was no measurable production function from which quantitative inferences about marginal productivities could be drawn; and such qualitative notions of a production function as the two managers possessed were mutually incompatible. To the public works administrator, a playground was a physical facility, serving as a green oasis in the crowded gray city. To the recreation administrator, a playground was a social facility, where children could play together with adult help and guidance.

How can human beings make rational decisions in circumstances like these? How are they to apply the marginal calculus? Or, if it does not apply, what do they substitute for it?

The phenomenon observed in Milwaukee is ubiquitous in human decision making. In organization theory it is usually referred to as *subgoal identification*. When the goals of an organization cannot be connected operationally with actions (when the production function can't be formulated in concrete terms), then decisions will be judged against subordinate goals that can be so connected. There is no unique determination of these subordinate goals. Their formulation will depend on the knowledge, experience, and organizational environment of the decision maker. In the face of this ambiguity, the formulation can also be influenced in subtle, and not so subtle, ways by his self-interest and power drives.

The phenomenon arises as frequently in individual as in social decision making and problem solving. Today, under the rubric of *problem representation*, it is a central research interest of cognitive psychology. Given a particular environment of stimuli, and a particular background of previous knowledge, how will a person organize this complex mass of information into a problem formulation that will facilitate his solution efforts? How did Newton's experience of the apple, if he had one, get represented as an instance of attraction of apple by Earth?

Phenomena like these provided the central theme for *Administrative Behavior*. That study represented "an attempt to construct tools useful in my own research in the field of public administration." The product was actually not so much a theory as prolegomena to a theory, stemming from the conviction "that decision making is the heart of administration, and that the vocabulary of administrative theory must be derived from the logic and psychology of human choice." It was, if you please, an exercise in problem representation.

On examination, the phenomenon of subgoal identification proved to be the visible tip of a very large iceberg. The shape of the iceberg is best appreciated by contrasting it with classical models of rational choice. The classical model calls for knowledge of all the alternatives that are open to choice. It calls for complete knowledge of, or ability to compute, the consequences that will follow on each of the alternatives. It calls for certainty in the decision maker's present and future evaluation of these consequences. It calls for the ability to compare consequences, no matter how diverse and heterogeneous, in terms of some consistent measure of utility. The task, then, was to replace the classical

model with one that would describe how decisions could be (and probably actually were) made when the alternatives of search had to be sought out, the consequences of choosing particular alternatives were only very imperfectly known both because of limited computational power and because of uncertainty in the external world, and the decision maker did not possess a general and consistent utility function for comparing heterogeneous alternatives.

Several procedures of rather general applicability and wide use have been discovered that transform intractable decision problems into tractable ones. One procedure already mentioned is to look for satisfactory choices instead of optimal ones. Another is to replace abstract, global goals with tangible subgoals, whose achievement can be observed and measured. A third is to divide up the decision-making task among many specialists, coordinating their work by means of a structure of communications and authority relations. All of these, and others, fit the general rubric of "bounded rationality," and it is now clear that the elaborate organizations that human beings have constructed in the modern world to carry out the work of production and government can only be understood as machinery for coping with the limits of man's abilities to comprehend and compute in the face of complexity and uncertainty.

This rather vague and general initial formulation of the idea of bounded rationality called for elaboration in two directions: greater formalization of the theory, and empirical verification of its main claims. During the decade that followed the publication of *Administrative Behavior*, substantial progress was made in both directions, some of it through the efforts of my colleagues and myself, much of it by other research groups that shared the same Zeitgeist.

B. *Empirical Studies*

The principal source of empirical data about organizational decision making has been straightforward "anthropological" field study, eliciting descriptions of decision-making procedures and observing the course of specific decision-making episodes. Examples are my study, with Guetzkow, Kozmetsky, and Tyndall (1954), of the ways in which accounting data were used in decision making in large corporations; and a series of studies, with Richard Cyert, James March, and others, of specific nonprogrammed policy decisions in a number of different companies (see Cyert, Simon, and Donald Trow). The latter line of work was greatly developed and expanded by Cyert and March and its theoretical implications for economics explored in their important work, *A Behavioral Theory of the Firm*.

At about the same time, the fortuitous availability of some data on businessmen's perceptions of a problem situation described in a business policy casebook enabled DeWitt Dearborn and me to demonstrate empirically the cognitive basis for identification with subgoals, the phenomenon that had so impressed me in the Milwaukee recreation study. The businessmen's perceptions of the principal problems facing the company described in the case were mostly determined by their own business experiences—sales and accounting executives identified a sales problem, manufacturing executives, a problem of internal organization.

Of course there is vastly more to be learned and tested about organizational decision making than can be dealt with in a handful of studies. Although many subsequent studies have been carried out in Europe and the United States, this domain is still grossly undercultivated (for references, see March, 1965; E. Johnsen, 1968; G. Eliasson, 1976). Among the reasons for the relative neglect of such studies, as contrasted, say, with laboratory experiments in social psychology, is that they are extremely costly and time consuming, with a high grist-to-grain ratio, the methodology for carrying them out is primitive, and satisfactory access to decision-making behavior is hard to secure. This part of economics has not yet acquired the habits of patience and persistence in the pursuit of facts that is exemplified in other domains by the work, say, of Simon Kuznets or of the architects of the MIT-SSRC-Penn econometric models.

C. *Theoretical Inquiries*

On the theoretical side, three questions seemed especially to call for clarification: what are the circumstances under which an employment relation will be preferred to some other form of contract as the arrangement for securing the performance of work; what is the relation between the classical theory of the firm and theories of organizational equilibrium first proposed by Chester Barnard; and what are the main characteristics of human rational choice in situations where complexity precludes omniscience?

The Employment Relation. A fundamental characteristic of modern industrial society is that most work is performed, not by individuals who produce products for sale, nor by individual contractors, but by persons who have accepted employment in a business firm and the authority relation with the employer that employment entails. Acceptance of authority means willingness to permit one's behavior to be determined by the employer, at least within some zone of indifference or acceptance. What is the advantage of this arrangement over a contract for specified goods or services? Why is so much of the world's work performed in large, hierarchic organizations?

Analysis showed (Simon, 1951) that a combination of two factors could account for preference for the employment contract over other forms of contracts: uncertainty as to which future behaviors would be advantageous to the employer, and a greater indifference of the employee as compared with the employer (within the former's area of acceptance) as to which of these behaviors he carried out. When the secretary is hired, the employer does not know what letters he will want her to type, and the secretary has no great preference for typing one letter rather than another. The employment contract permits the choice to be postponed until the uncertainty is resolved, with little cost to the employee and great advantage to the employer. The explanation is closely analogous to one Jacob Marschak had proposed for liquidity preference. Under conditions of uncertainty it is advantageous to hold resources in liquid, flexible form.

Organizational Equilibrium. Barnard had described the survival of organizations in terms of the motivations that make their participants (employees, investors, customers, suppliers) willing to remain in the system. In *Administrative Behavior*, I had developed this notion further into a motivational theory of the balance between the inducements that were provided by organizations to their participants, and the contributions those participants made to the organizations' resources.

A formalization of this theory (Simon, 1952; 1953) showed its close affinity to the classical theory of the firm, but with an important and instructive difference. In comparing the two theories, each inducement-contribution relation became a supply schedule for the firm. The survival conditions became the conditions for positive profit. But while the classical theory of the firm assumes that all profits accrue to a particular set of participants, the owners, the organization theory treats the surplus more symmetrically, and does not predict how it will be distributed. Hence the latter theory leaves room, under conditions of monopoly and imperfect competition, for bargaining among the participants (for example, between labor and owners) for the surplus. The survival conditions—positive profits rather than maximum profits—also permit a departure from the assumptions of perfect rationality.

Mechanisms of Bounded Rationality. In *Administrative Behavior*, bounded rationality is largely characterized as a residual category—rationality is bounded when it falls short of omniscience. And the failures of omniscience are largely failures of knowing all the alternatives, uncertainty about relevant exogenous events, and inability to calculate consequences. There was needed a more positive and formal characterization of the mechanisms of choice under conditions of bounded rationality. Two papers (Simon, 1955; 1956a) undertook first steps in that direction.

Two concepts are central to the characterization: *search* and *satisficing*. If the alternatives for choice are not given initially to the decision maker, then he must search for them. Hence, a theory of bounded rationality must incorporate a theory of search. This idea was

later developed independently by George Stigler in a very influential paper that took as its example of a decision situation the purchase of a second-hand automobile. Stigler poured the search theory back into the old bottle of classical utility maximization, the cost of search being equated with its marginal return. In my 1956 paper, I had demonstrated the same formal equivalence, using as my example a dynamic programming formulation of the process of selling a house.

But utility maximization, as I showed, was not essential to the search scheme—fortunately, for it would have required the decision maker to be able to estimate the marginal costs and returns of search in a decision situation that was already too complex for the exercise of global rationality. As an alternative, one could postulate that the decision maker had formed some *aspiration* as to how good an alternative he should find. As soon as he discovered an alternative for choice meeting his level of aspiration, he would terminate the search and choose that alternative. I called this mode of selection *satisficing*. It had its roots in the empirically based psychological theories, due to Lewin and others, of aspiration levels. As psychological inquiry had shown, aspiration levels are not static, but tend to rise and fall in consonance with changing experiences. In a benign environment that provides many good alternatives, aspirations rise; in a harsher enviornment, they fall.

In long-run equilibrium it might even be the case that choice with dynamically adapting aspiration levels would be equivalent to optimal choice, taking the costs of search into account. But the important thing about the search and satisficing theory is that it showed how choice could actually be made with reasonable amounts of calculation, and using very incomplete information, without the need of performing the impossible—of carrying out this optimizing procedure.

D. *Summary*

Thus, by the middle 1950's, a theory of bounded rationality had been proposed as an alternative to classical omniscient rationality, a significant number of empirical studies had been carried out that showed actual business decision making to conform reasonably well with the assumptions of bounded rationality but not with the assumptions of perfect rationality, and key components of the theory—the nature of the authority and employment relations, organizational equilibrium, and the mechanisms of search and satisficing—had been elucidated formally. In the remaining parts of this paper, I should like to trace subsequent developments of decision-making theory, including developments competitive with the theory of bounded rationality, and then to comment on the implications (and potential implications) of the new descriptive theory of decision for political economy.

III. The Neoclassical Revival

Peering forward from the late 1950's, it would not have been unreasonable to predict that theories of bounded rationality would soon find a large place in the mainstream of economic thought. Substantial progress had been made in providing the theories with some formal structure, and an increasing body of empirical evidence showed them to provide a far more veridical picture of decision making in business organizations than did the classical concepts of perfect rationality.

History has not followed any such simple course, even though many aspects of the Zeitgeist were favorable to movement in this direction. During and after World War II, a large number of academic economists were exposed directly to business life, and had more or less extensive opportunities to observe how decisions were actually made in business organizations. Moreover, those who became active in the development of the new management science were faced with the necessity of developing decision-making procedures that could actually be applied in practical situations. Surely these trends would be conducive to moving the basic assumptions of economic rationality in the direction of greater realism.

But these were not the only things that were happening in economics in the postwar

period. First, there was a vigorous reaction that sought to defend classical theory from behavioralism on methodological grounds. I have already commented on these methodological arguments in the first part of my talk. However deeply one may disagree with them, they were stated persuasively and are still influential among academic economists.

Second, the rapid spread of mathematical knowledge and competence in the economics profession permitted the classical theory, especially when combined with statistical decision theory and the theory of games due to von Neumann and Morgenstern, to develop to new heights of sophistication and elegance, and to expand to embrace, albeit in highly stylized form, some of the phenomena of uncertainty and imperfect information. The flowering of mathematical economics and econometrics has provided two generations of economic theorists with a vast garden of formal and technical problems that have absorbed their energies and postponed encounters with the inelegancies of the real world.

If I sound mildly critical of these developments, I should confess that I have also been a part of them, admire them, and would be decidedly unhappy to return to the premathematical world they have replaced. My concern is that the economics profession has exhibited some of the serial one-thing-at-a-time character of human rationality, and has seemed sometimes to be unable to distribute its attention in a balanced fashion among neoclassical theory, macroeconometrics, and descriptive decision theory. As a result, not as much professional effort has been devoted to the latter two, and especially the third, as one might have hoped and expected. The Heartland is more overpopulated than ever, while rich lands in other parts of the empire go untilled.

A. *Search and Information Transfer*

Let me allude to just three of the ways in which classical theory has sought to cope with some of its traditional limitations, and has even sought to make the development of a behavioral theory, incorporating psychological assumptions, unnecessary. The first was to introduce search and information transfer explicitly as economic activities, with associated costs and outputs, that could be inserted into the classical production function. I have already referred to Stigler's 1961 paper on the economics of information, and my own venture in the same direction in the 1956 essay cited earlier.

In theory of this genre, the decision maker is still an individual. A very important new direction, in which decisions are made by groups of individuals, in teams or organizations, is the economic theory of teams developed by Jacob Marschak and Roy Radner. Here we see genuine organizational phenomena—specialization of decision making as a consequence of the costs of transmitting information—emerge from the rational calculus. Because the mathematical difficulties are formidable, the theory remains largely illustrative and limited to very simple situations in miniature organizations. Nevertheless, it has greatly broadened our understanding of the economics of information.

In none of these theories—any more than in statistical decision theory or the theory of games—is the assumption of perfect maximization abandoned. Limits and costs of information are introduced, not as psychological characteristics of the decision maker, but as part of his technological environment. Hence, the new theories do nothing to alleviate the computational complexities facing the decision maker—do not see him coping with them by heroic approximation, simplifying and satisficing, but simply magnify and multiply them. Now he needs to compute not merely the shapes of his supply and demand curves, but, in addition, the costs and benefits of computing those shapes to greater accuracy as well. Hence, to some extent, the impression that these new theories deal with the hitherto ignored phenomena of uncertainty and information transmission is illusory. For many economists, however, the illusion has been persuasive.

B. *Rational Expectations Theory*

A second development in neoclassical theory on which I wish to comment is the so-called "rational expectations" theory.

There is a bit of historical irony surrounding its origins. I have already described the management science inquiry of Holt, Modigliani, Muth, and myself that developed a dynamic programming algorithm for the special (and easily computed) case of quadratic cost functions. In this case, the decision rules are linear, and the probability distributions of future events can be replaced by their expected values, which serve as certainty equivalents (see Simon, 1956; Henri Theil, 1957).

Muth imaginatively saw in this special case a paradigm for rational behavior under uncertainty. What to some of us in the HMMS research team was an approximating, satisficing simplification, served for him as a major line of defense for perfect rationality. He said in his seminal 1961 *Econometrica* article, "It is sometimes argued that the assumption of rationality in economics leads to theories inconsistent with, or inadequate to explain, observed phenomena, especially changes over time... Our hypothesis is based on exactly the opposite point of view: that dynamic economic models do not assume enough rationality" (p. 316).

The new increment of rationality that Muth proposed was that "expectations, since they are informed predictions of future events, are essentially the same as the predictions of the relevant economic theory" (p. 316). He would cut the Gordian knot. Instead of dealing with uncertainty by elaborating the model of the decision process, he would once and for all—if his hypothesis were correct—make process irrelevant. The subsequent vigorous development of rational expectations theory, in the hands of Sargent, Lucas, Prescott, and others, is well known to most readers (see, for example, Lucas, 1975).

It is too early to render a final verdict on the rational expectations theory. The issue will ultimately be decided, as all scientific debates should be, by a gradual winnowing of the empirical evidence, and that winnowing process has just begun. Meanwhile, certain grave theoretical difficulties have already been noticed. As Muth himself has pointed out, it is rational (i.e., profit maximizing) to use the "rational expectations" decision rule if the relevant cost equations are in fact quadratic. I have suggested elsewhere (1978a) that it might therefore be less misleading to call the rule a "consistent expectations" rule.

Perhaps even more important, Albert Ando and Benjamin Friedman (1978, 1979) have shown that the policy implications of the rational expectations rule are quite different under conditions where new information continually becomes available to the system, structural changes occur, and the decision maker learns, than they are under steady-state conditions. For example, under the more dynamic conditions, monetary neutrality—which in general holds for the static consistent expectations models—is no longer guaranteed for any finite time horizon.

In the recent "revisionist" versions of consistent expectations theory, moreover, where account is taken of a changing environment of information, various behavioral assumptions reappear to explain how expectations are formed—what information decision makers will consider, and what they will ignore. But unless these assumptions are to be made on a wholly *ad hoc* and arbitrary basis, they create again the need for an explicit and valid theory of the decision-making *process* (see Simon, 1958a; B. Friedman, 1979).

C. *Statistical Decision Theory and Game Theory*

Statistical decision theory and game theory are two other important components of the neoclassical revival. The former addresses itself to the question of incorporating uncertainty (or more properly, risk) into the decision-making models. It requires heroic assumptions about the information the decision maker has concerning the probability distributions of the relevant variables, and simply increases by orders of magnitude the computational problems he faces.

Game theory addresses itself to the "outguessing" problem that arises whenever an economic actor takes into account the possible reactions to his own decisions of the other actors. To my mind, the main product of the very elegant apparatus of game theory has been to demonstrate quite clearly that it is virtually impossible to define an unambiguous

criterion of rationality for this class of situations (or, what amounts to the same thing, a definitive definition of the "solution" of a game). Hence, game theory has not brought to the theories of oligopoly and imperfect competition the relief from their contradictions and complexities that was originally hoped for it. Rather, it has shown that these difficulties are ineradicable. We may be able to reach consensus that a certain criterion of rationality is appropriate to a particular game, but if someone challenges the consensus, preferring a different criterion, we will have no logical basis for persuading him that he is wrong.

D. *Conclusion*

Perhaps I have said enough about the neoclassical revival to suggest why it has been a highly attractive commodity in competition with the behavioral theories. To some economists at least, it has held open the possibility and hope that important questions that had been troublesome for classical economics could now be addressed without sacrifice of the central assumption of perfect rationality, and hence also with a maximum of a priori inference and a minimum of tiresome grubbing with empirical data. I have perhaps said enough also with respect to the limitations of these new constructs to indicate why I do not believe that they solve the problems that motivated their development.

IV. Advances in the Behavioral Theory

Although they have played a muted role in the total economic research activity during the past two decades, theories of bounded rationality and the behavioral theory of the business firm have undergone steady development during that period. Since surveying the whole body of work would be a major undertaking, I shall have to be satisfied here with suggesting the flavor of the whole by citing a few samples of different kinds of important research falling in this domain. Where surveys on particular topics have been published, I will limit myself to references to them.

First, there has been work in the psychological laboratory and the field to test whether people in relatively simple choice situations behave as statistical decision theory (maximization of expected utilities) say they do. Second, there has been extensive psychological research, in which Allen Newell and I have been heavily involved, to discover the actual microprocesses of human decision making and problem solving. Third, there have been numerous empirical observations—most of them in the form of "case studies"—of the actual processes of decision making in organizational and business contexts. Fourth, there have been reformulations and extensions of the theory of the firm replacing classical maximization with behavioral decision postulates.

A. *Utility Theory and Human Choice*

The axiomatization of utility and probability after World War II and the revival of Bayesian statistics opened the way to testing empirically whether people behaved in choice situations so as to maximize subjective expected utility (*SEU*). In early studies, using extremely simple choice situations, it appeared that perhaps they did. When even small complications were introduced into the situations, wide departures of behavior from the predictions of *SEU* theory soon became evident. Some of the most dramatic and convincing empirical refutations of the theory have been reported by D. Kahneman and A. Tversky, who showed that under one set of circumstances, decision makers gave far too little weight to prior knowledge and based their choices almost entirely on new evidence, while in other circumstances new evidence had little influence on opinions already formed. Equally large and striking departures from the behavior predicted by the *SEU* theories were found by Howard Kunreuther and his colleagues in their studies of individual decisions to purchase or not to purchase flood insurance. On the basis of these and other pieces of evidence, the conclusion seems unavoidable that the *SEU* theory does not provide a good prediction—not even a good approximation—of actual behavior.

Notice that the refutation of the theory has to do with the *substance* of the decisions, and not just the process by which they are reached. It is not that people do not go through the calculations that would be required to reach the *SEU* decision—neoclassical thought has never claimed that they did. What has been shown is that they do not even behave *as if* they had carried out those calculations, and that result is a direct refutation of the neoclassical assumptions.

B. *Psychology of Problem Solving*

The evidence on rational decision making is largely negative evidence, evidence of what people do *not* do. In the past twenty years a large body of positive evidence has also accumulated about the processes that people use to make difficult decisions and solve complex problems. The body of theory that has been built up around this evidence is called information processing psychology, and is usually expressed formally in computer programming languages. Newell and I have summed up our own version of this theory in our book, *Human Problem Solving*, which is part of a large and rapidly growing literature that assumes an information processing framework and makes use of computer simulation as a central tool for expressing and testing theories.

Information processing theories envisage problem solving as involving very selective search through problem spaces that are often immense. Selectivity, based on rules of thumb or "heuristics," tends to guide the search into promising regions, so that solutions will generally be found after search of only a tiny part of the total space. Satisficing criteria terminate search when satisfactory problem solutions have been found. Thus, these theories of problem solving clearly fit within the framework of bounded rationality that I have been expounding here.

By now the empirical evidence for this general picture of the problem solving process is extensive. Most of the evidence pertains to relatively simple, puzzle-like situations of the sort that can be brought into the psychological laboratory for controlled study, but a great deal has been learned, also, about professional level human tasks like making medical diagnoses, investing in portfolios of stocks and bonds, and playing chess. In tasks of these kinds, the general search mechanisms operate in a rich context of information stored in human long-term memory, but the general organization of the process is substantially the same as for the simpler, more specific tasks.

At the present time, research in information processing psychology is proceeding in several directions. Exploration of professional level skills continues. A good deal of effort is now being devoted also to determining how initial representations for new problems are acquired. Even in simple problem domains, the problem solver has much latitude in the way he formulates the problem space in which he will search, a finding that underlines again how far the actual process is from a search for a uniquely determined optimum (see J. R. Hayes and Simon).

The main import for economic theory of the research in information processing psychology is to provide rather conclusive empirical evidence that the decision-making process in problem situations conforms closely to the models of bounded rationality described earlier. This finding implies, in turn, that choice is not determined uniquely by the objective characteristics of the problem situation, but depends also on the particular heuristic process that is used to reach the decision. It would appear, therefore, that a model of process is an essential component in any positive theory of decision making that purports to describe the real world, and that the neoclassical ambition of avoiding the necessity for such a model is unrealizable (Simon, 1978a).

C. *Organizational Decision Making*

It would be desirable to have, in addition to the evidence from the psychological research just described, empirical studies of the process of decision making in organizational contexts. The studies of individual problem solving and decision making do not touch on the many social-psychological factors that enter into the decision process in organiza-

tions. A substantial number of investigations have been carried out in the past twenty years of the decision-making process in organizations, but they are not easily summarized. The difficulty is that most of these investigations have taken the form of case studies of specific decisions or particular classes of decisions in individual organizations. To the best of my knowledge, no good review of this literature has been published, so that it is difficult even to locate and identify the studies that have been carried out.[3] Nor have any systematic methods been developed and tested for distilling out from these individual case studies their implications for the general theory of the decision-making process.

The case studies of organizational decision making, therefore, represent the natural history stage of scientific inquiry. They provide us with a multitude of facts about the decision-making process—facts that are almost uniformly consistent with the kind of behavioral model that has been proposed here. But we do not yet know how to use these facts to test the model in any formal way. Nor do we quite know what to do with the observation that the specific decision-making procedures used by organizations differ from one organization to another, and within each organization, even from one situation to another. We must not expect from these data generalizations as neat and precise as those incorporated in neoclassical theory.

Perhaps the closest approach to a method for extracting theoretically relevant information from case studies is computer simulation. By converting empirical evidence about a decision-making process into a computer program, a path is opened both for testing the adequacy of the program mechanisms for explaining the data, and for discovering the key features of the program that account, qualitatively, for the interesting and important characteristics of its behavior. Examples of the use of this technique are G.P.E. Clarkson's simulation of the decision making of an investment trust officer, Cyert, E. A. Feigenbaum, and March's simulation of the history of a duopoly, and C. P. Bonini's model of the effects of accounting information and supervisory pressures in altering employee motivations in a business firm. The simulation methodology is discussed from a variety of viewpoints in Dutton and Starbuck.[4]

D. *Theories of the Business Firm*

The general features of bounded rationality—selective search, satisficing, and so on—have been taken as the starting points for a number of attempts to build theories of the business firm incorporating behavioral assumptions. Examples of such theories would include the theory of Cyert and March, already mentioned; William Baumol's theory of sales maximization subject to minimum profit constraints; Robin Marris' models of firms whose goals are stated in terms of rates of growth; Harvey Leibenstein's theory of "X-inefficiency" that depresses production below the theoretically attainable; Janos Kornai's dichotomy between supply-driven and demand-driven management; Oliver Williamson's theory of transactional costs; the evolutionary models of Richard Nelson and Sidney Winter (1973); Cyert and Morris DeGroot's (1974) models incorporating adaptive learning; Radner's (1975a,b) explicit satisficing models; and others.

Characterized in this way, there seems to be little commonality among all of these theories and models, except that they depart in one way or another from the classical assumption of perfect rationality in firm decision making. A closer look, however, and a more abstract description of their assumptions, shows that they share several basic characteristics. Most of them depart from the assumption of profit maximization in the short run, and replace it with an assumption

[3] For leads into the literature, see March and Simon; March; Johnsen; J. M. Dutton and W. H Starbuck. However, there are large numbers of specific case studies, some of them carried out as thesis projects, some concerned with particular fields of business application, which have never been recorded in these reference sources (for example, Eliasson, 1976).

[4] In addition to simulations of the firm, there are very interesting and potentially important efforts to use simulation to build bridges directly from decision theory to political economy. See G. Orcutt and R. Caldwell-Wertheimer, and Eliasson (1978).

of goals defined in terms of targets—that is, they are to greater or lesser degree satisficing theories. If they do retain maximizing assumptions, they contain some kind of mechanism that prevents the maximum from being attained, at least in the short run. In the Cyert-March theory, and that of Leibenstein, this mechanism can be viewed as producing "organizational slack," the magnitude of which may itself be a function of motivational and environmental variables.

Finally, a number of these theories assume that organizational learning takes place, so that if the environment were stationary for a sufficient length of time, the system equilibrium would approach closer and closer to the classical profit-maximizing equilibrium. Of course they generally also assume that the environmental disturbances will generally be large enough to prevent the classical solution from being an adequate approximation to the actual behavior.

The presence of something like organizational slack in a model of the business firm introduces complexity in the firm's behavior in the short run. Since the firm may operate very far from any optimum, the slack serves as a buffer between the environment and the firm's decisions. Responses to environmental events can no longer be predicted simply by analyzing the "requirements of the situation," but depend on the specific decision processes that the firm employs. However well this characteristic of a business firm model corresponds to reality, it reduces the attractiveness of the model for many economists, who are reluctant to give up the process-independent predictions of classical theory, and who do not feel at home with the kind of empirical investigation that is required for disclosing actual real world decision processes.

But there is another side to the matter. If, in the face of identical environmental conditions, different decision mechanisms can produce different firm behaviors, this sensitivity of outcomes to process can have important consequences for analysis at the level of markets and the economy. Political economy, whether descriptive or normative, cannot remain indifferent to this source of variability in response. At the very least it demands that—before we draw policy conclusions from our theories, and particularly before we act on those policy conclusions—we carry out sensitivity analyses to test how far our conclusions would be changed if we made different assumptions about the decision mechanisms at the micro level.

If our conclusions are robust—if they are not changed materially by substituting one or another variant of the behavioral model for the classical model—we will gain confidence in our predictions and recommendations; if the conclusions are sensitive to such substitutions, we will use them warily until we can determine which micro theory is the correct one.

As reference to the literature cited earlier in this section will verify, our predictions of the operations of markets and of the economy *are* sensitive to our assumptions about mechanisms at the level of decision processes. Moreover, the assumptions of the behavioral theories are almost certainly closer to reality than those of the classical theory. These two facts, in combination, constitute a direct refutation of the argument that the unrealism of the assumptions of the classical theory is harmless. We cannot use the *in vacua* version of the law of falling bodies to predict the sinking of a heavy body in molasses. The predictions of the classical and neoclassical theories and the policy recommendations derived from them must be treated with the greatest caution.

V. Conclusion

There is a saying in politics that "you can't beat something with nothing." You can't defeat a measure or a candidate simply by pointing to defects and inadequacies. You must offer an alternative.

The same principle applies to scientific theory. Once a theory is well entrenched, it will survive many assaults of empirical evidence that purports to refute it unless an alternative theory, consistent with the evidence, stands ready to replace it. Such conservative protectiveness of established beliefs is, indeed, not unreasonable. In the first place, in empirical science we aspire only to approxi-

mate truths; we are under no illusion that we can find a single formula, or even a moderately complex one, that captures the whole truth and nothing else. We are committed to a strategy of successive approximations, and when we find discrepancies between theory and data, our first impulse is to patch rather than to rebuild from the foundations.

In the second place, when discrepancies appear, it is seldom immediately obvious where the trouble lies. It may be located in the fundamental assumptions of the theory, but it may as well be merely a defect in the auxiliary hypotheses and measurement postulates we have had to assume in order to connect theory with observations. Revisions in these latter parts of the structure may be sufficient to save the remainder.

What then is the present status of the classical theory of the firm? There can no longer be any doubt that the micro assumptions of the theory—the assumptions of perfect rationality—are contrary to fact. It is not a question of approximation; they do not even remotely describe the processes that human beings use for making decisions in complex situations.

Moreover, there is an alternative. If anything, there is an embarrassing richness of alternatives. Today, we have a large mass of descriptive data, from both laboratory and field, that show how human problem solving and decision making actually take place in a wide variety of situations. A number of theories have been constructed to account for these data, and while these theories certainly do not yet constitute a single coherent whole, there is much in common among them. In one way or another, they incorporate the notions of bounded rationality: the need to search for decision alternatives, the replacement of optimization by targets and satisficing goals, and mechanisms of learning and adaptation. If our interest lies in descriptive decision theory (or even normative decision theory), it is now entirely clear that the classical and neoclassical theories have been replaced by a superior alternative that provides us with a much closer approximation to what is actually going on.

But what if our interest lies primarily in normative political economy rather than in the more remote regions of the economic sciences? Is there then any reason why we should give up the familiar theories? Have the newer concepts of decision making and the firm shown their superiority "for purposes of economic analysis"?

If the classical and neoclassical theories were, as is sometimes argued, simply powerful tools for deriving aggregative consequences that held alike for both perfect and bounded rationality, we would have every reason to retain them for this purpose. But we have seen, on the contrary, that neoclassical theory does not always lead to the same conclusions at the level of aggregate phenomena and policy as are implied by the postulate of bounded rationality, in any of its variants. Hence, we cannot defend an uncritical use of these contrary-to-fact assumptions by the argument that their veridicality is unimportant. In many cases, in fact, this veridicality may be crucial to reaching correct conclusions about the central questions of political economy. Only a comparison of predictions can tell us whether a case before us is one of these.

The social sciences have been accustomed to look for models in the most spectacular successes of the natural sciences. There is no harm in that, provided that it is not done in a spirit of slavish imitation. In economics, it has been common enough to admire Newtonian mechanics (or, as we have seen, the Law of Falling Bodies), and to search for the economic equivalent of the laws of motion. But this is not the only model for a science, and it seems, indeed, not to be the right one for our purposes.

Human behavior, even rational human behavior, is not to be accounted for by a handful of invariants. It is certainly not to be accounted for by assuming perfect adaptation to the environment. Its basic mechanisms may be relatively simple, and I believe they are, but that simplicity operates in interaction with extremely complex boundary conditions imposed by the environment and by the very facts of human long-term memory and of the capacity of human beings, individually and collectively, to learn.

If we wish to be guided by a natural science metaphor, I suggest one drawn from biology

rather than physics (see Newell and Simon, 1976). Obvious lessons are to be learned from evolutionary biology, and rather less obvious ones from molecular biology. From molecular biology, in particular, we can glimpse a picture of how a few basic mechanisms—the DNA of the Double Helix, for example, or the energy transfer mechanisms elucidated so elegantly by Peter Mitchell—can account for a wide range of complex phenomena. We can see the role in science of laws of qualitative structure, and the power of qualitative as well as quantitative explanation.

I am always reluctant to end a talk about the sciences of man in the future tense. It conveys too much the impression that these are potential sciences which may some day be actualized, but that do not really exist at the present time. Of course that is not the case at all. However much our knowledge of human behavior falls short of our need for such knowledge, still it is enormous. Sometimes we tend to discount it because so many of the phenomena are accessible to us in the very activity of living as human beings among human beings that it seems commonplace to us. Moreover, it does not always answer the questions for which we need answers. We cannot predict very well the course of the business cycle nor manage the employment rate. (We cannot, it might be added, predict very well the time of the next thunderstorm in Stockholm, or manage the earth's climates.)

With all these qualifications and reservations, we do understand today many of the mechanisms of human rational choice. We do know how the information processing system called Man, faced with complexity beyond his ken, uses his information processing capacities to seek out alternatives, to calculate consequences, to resolve uncertainties, and thereby—sometimes, not always—to find ways of action that are sufficient unto the day, that satisfice.

REFERENCES

A. A. Alchian, "Uncertainty, Evolution, and Economic Theory," *J. Polit. Econ.*, June 1950, *58*, 211-21.

A. Ando, "On a Theoretical and Empirical Basis of Macroeconometric Models," paper presented to the NSF-NBER Conference on Macroeconomic Modeling, Ann Arbor, Oct. 1978.

Chester I. Barnard, *The Functions of the Executive*, Cambridge, Mass. 1938.

William Baumol, *Business Behavior, Value and Growth*, New York 1959.

G. S. Becker, "Irrational Behavior and Economic Theory," *J. Polit. Econ.*, Feb. 1962, *70*, 1–13.

Charles P. Bonini, *Simulation of Information and Decision Systems in the Firm*, Englewood Cliffs 1963.

Alfred Chandler, *Strategy and Structure*, Cambridge, Mass. 1962.

N. C. Churchill, W. W. Cooper, and T. Sainsbury, "Laboratory and Field Studies of the Behavioral Effects of Audits," in C. P. Bonini et al., eds., *Management Controls*, New York 1964.

G. P. E. Clarkson, "A Model of the Trust Investment Process," in E. A. Feigenbaum and J. Feldman, eds., *Computers and Thought*, New York 1963.

John R. Commons, *Institutional Economics*, Madison 1934.

R. M. Cyert, E. A. Feigenbaum, and J. G. March, "Models in a Behavioral Theory of the Firm," *Behav. Sci.*, Apr. 1959, *4*, 81–95.

———**and M. H. DeGroot**, "Rational Expectations and Bayesian Analysis," *J. Polit. Econ.*, May/June 1974, *82*, 521–36.

——— **and** ——— "Adaptive Utility," in R. H. Day and T. Groves, eds., *Adaptive Economic Models*, New York 1975, 233–46.

———**and James G. March**, *A Behavioral Theory of the Firm*, Englewood Cliffs 1963.

——— **and H. A. Simon**, "Theory of the Firm: Behavioralism and Marginalism," unpublished work. paper, Carnegie-Mellon Univ. 1971.

———, ———, **and D. B. Trow**, "Observation of a Business Decision," *J. Bus., Univ. Chicago*, Oct. 1956, *29*, 237–48.

D. C. Dearborn and H. A. Simon, "Selective Perception: The Identifications of Executives," *Sociometry*, 1958, *21*, 140–144; reprinted in *Administrative Behavior*, ch. 15, 3d ed., New York 1976.

J. M. Dutton and W. H. Starbuck, *Computer Simulation of Human Behavior*, New York 1971.

G. Eliasson, *Business Economic Planning*, New York 1976.

———, *A Micro-to-Macro Model of the Swedish Economy*, Stockholm 1978.

B. M. Friedman, "Optimal Expectations and the Extreme Information Assumptions of 'Rational Expectations' Macromodels," *J. Monet. Econ.*, Jan. 1979 *5*, 23–41.

———, "A Discussion of the Methodological Premises of Professors Lucas and Sargent," in *After the Phillips Curve: The Persistence of High Inflation and High Unemployment*, Boston 1978.

Milton Friedman, *Essays in Positive Economics*, Chicago 1953.

J. R. Hayes and H. A. Simon, "Understanding Written Problem Instructions," in W. Gregg, ed., *Knowledge and Cognition*, Potomac 1974, 167–200.

A. O. Hirschman, *Exit, Voice and Loyalty*, Cambridge, Mass. 1970.

Charles C. Holt, Franco Modigliani, John F. Muth, and Herbert A. Simon, *Planning Production, Inventories and Work Force*, Englewood Cliffs 1960.

Y. Ijiri and H. A. Simon, *Skew Distributions and the Sizes of Business Firms*, Amsterdam 1977.

E. Johnsen, *Studies in Multiobjective Decision Models*, Lund 1968.

D. W. Jorgenson and C. D. Siebert, "A Comparison of Alternative Theories of Corporate Investment Behavior," *Amer. Econ. Rev.*, Sept. 1968, *58*, 681–712.

D. Kahneman and A. Tversky, "On the Psychology of Prediction," *Psychol. Rev.*, July 1973, *80*, 237–51.

Janos Kornai, *Anti-Equilibrium*, Amsterdam 1971.

Howard Kunreuther et al., *Disaster Insurance Protection: Public Policy Lessons*, New York 1978.

Harvey Leibenstein, *Beyond Economic Man*, Cambridge, Mass. 1976.

J. Lesourne, *A Theory of the Individual for Economic Analysis*, Vol. 1, Amsterdam 1977.

R. E. Lucas, Jr., "An Equilibrium Model of the Business Cycle," *J. Polit. Econ.*, Dec. 1975, *83*, 1113–44.

———, "On the Size Distribution of Business Firms," *Bell J. Econ.*, Autumn 1978, *9*, 508–23.

James G. March, *Handbook of Organizations*, Chicago 1965.

——— **and H. A. Simon**, *Organizations*, New York 1958.

Robin Marris, *The Economic Theory of "Managerial" Capitalism*, London 1964.

Jacob Marschak, "Role of Liquidity under Complete and Incomplete Information," *Amer. Econ. Rev. Proc.*, May 1949, *39*, 182–95.

——— **and Roy Radner**, *Economic Theory of Teams*, New Haven 1972.

Alfred Marshall, *Principles of Economics*, 8th ed., New York 1920.

E. S. Mason, "Comment," in Bernard T. Haley, ed., *A Survey of Contemporary Economics*, Vol. II, Homewood 1952, 221–22.

J. M. Montias, *The Structure of Economic Systems*, New Haven 1976.

J. F. Muth, "Rational Expectations and the Theory of Price Movements," *Econometrica*, July 1961, *29*, 315–53.

———, "Optimal Properties of Exponentially Weighted Forecasts," *J. Amer. Statist. Assn.*, June 1960, *55*, 299–306.

R. R. Nelson, and S. Winter, "Toward an Evolutionary Theory of Economic Capabilities," *Amer. Econ. Rev. Proc.*, May 1973, *63*, 440–49.

——— **and** ———, "Neoclassical vs. Evolutionary Theories of Economic Growth," *Econ. J.*, Dec. 1974, *84*, 886–905.

Allen Newell and Herbert A. Simon, *Human Problem Solving*, Englewood Cliffs, 1972.

——— **and** ———, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM*, Mar. 1976, *19*,113–26.

G. Orcutt, and R. Caldwells-Wertheimer II, *Policy Exploration through Microanalytic Simulation*, Washington 1976.

A. Papandreou, "Some Basic Problems in the Theory of the Firm," in Bernard F. Haley, ed., *A Survey of Contemporary Economics*, Vol. II, Homewood 1952.

E. H. **Phelps-Brown**, "The Meaning of the Fitted Cobb-Douglas Function," *Quart. J. Econ.*, Nov. 1957, *71*, 546–60.

R. **Radner**, (1975a) "A Behavioral Model of Cost Reduction," *Bell J. Econ.*, Spring 1975, *6*, 196–215.

———, (1975b) "Satisficing," *J. Math. Econ.*, June–Sept. 1975, *2*, 253–62.

David R. Roberts, *Executive Compensation*, Glencoe 1959.

P. A. Samuelson, "Discussion: Problems of Methodology," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 231–36.

Henry Schultz, *The Theory and Measurement of Demand*, Chicago 1938.

Herbert A. Simon, *Administrative Behavior*, New York 1947; 3d ed. 1976.

———, "A Formal Theory of the Employment Relation," *Econometrica*, July 1951, *19*, 293–305

———, "A Comparison of Organization Theories," *Rev. Econ. Stud.*, No. 1, 1952, *20*, 40–48.

———, "A Behavioral Model of Rational Choice," *Quart. J. Econ.*, Feb. 1955, *69*, 99–118.

———, "Rational Choice and the Structure of the Environment," *Psychol. Rev.*, Mar. 1956, *63*, 129–38.

———, "Dynamic Programming under Uncertainty with a Quadratic Criterion Function," *Econometrica*, Jan. 1956, *24*, 74–81.

———, *Models of Man*, New York 1957.

———, "The Compensation of Executives," *Sociometry*, 1957, *20*, 32–35.

———, "Theories of Decision Making in Economics and Behavioral Science," *Amer. Econ. Rev.*, June 1959, *49*, 223–83.

———, "Discussion: Problems of Methodology," *Amer. Econ. Rev. Proc.*, May 1963, *53*, 229–31.

———, "From Substantive to Procedural Rationality," in Spiro J. Latsis, ed., *Methodological Appraisal in Economics*, Cambridge 1976.

———, (1978a) "Rationality as Process and as Product of Thought," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 1–16.

———, (1978b) "On How to Decide What to Do," *Bell J. Econ.*, Autumn 1978, *9*, 494–507.

———, **G. Kozmetsky, H. Guetzkow, and G. Tyndall**, *Centralization vs. Decentralization in Organizing the Controller's Department*, New York 1954; reprinted Houston 1978.

——— **and F. K. Levy**, "A Note on the Cobb-Douglas Function," *Rev. Econ. Stud.*, June 1963, *30*, 93–94.

G. J. Stigler, "The Economics of Information," *J. Polit. Econ.*, June 1961, *69*, 213–15.

H. Theil, "A Note on Certainty Equivalence in Dynamic Planning," *Econometrica*, Apr. 1957, *25*, 346–49.

John von Neumann and Oscar Morgenstern, *Theory of Games and Economic Behavior*, Princeton 1944.

A. A. Walters, "Production and Cost Functions: An Econometric Survey," *Econometrica*, Jan.–Apr. 1963, *31*, 1–66.

Oliver Williamson, *Markets and Hierarchies: Analysis and Antitrust Implications*, New York 1975.

S. Winter, "Satisficing, Selection, and the Innovating Remnant," *Quart. J. Econ.*, May 1971, *85*, 237–61.

The Output Distribution Frontier: Alternatives to Income Taxes and Transfers for Strong Equality Goals

By William J. Baumol and Dietrich Fischer*

It has long been suggested that moves toward equality can have serious disincentive effects and, after a point, lead to unacceptable losses in real income.[1] We will now offer some theoretical grounds for believing that an approximation to equality achieved via the traditional instruments—transfer payments and progressive taxation—will cause an income loss far more serious than many of us have realized. We will prove that under a set of reasonable assumptions, any attempt to guarantee *absolute equality* of incomes using only progressive income taxes and transfers for the purpose *must, at least in theory, reduce society's output to zero*! However, we do *not* conclude from this that the search for very much increased equality is quixotic. Rather, we take this as a criticism of the means so far used for the purpose. What is called for is an exercise in imagination and ingenuity, to find some alternative ways to go about this quest. We will then show explicitly and examine an alternative procedure that, at least in theory, can achieve *any* desired degree of equality without necessarily exacting a serious loss in output, and will end by discussing briefly the possibility of practical approximations to such an arrangement.

I. The High Output Cost of Equalization via Progressive Taxes and Transfers

This section will show that, at least in theory, the use of progressive taxes and transfers to guarantee equality must incur a startlingly great loss in the economy's output. For this purpose, we must first examine the nature of the individual's input supply choice—the tradeoff between income y_i, and unused input (leisure) $v_i = r_i^* - r_i$, where r_i^* is the total input quantity she has available, and r_i is the amount she actually supplies (her labor time). Figure 1 depicts a linear budget locus Wr^* representing individual i's tradeoff between leisure v_i, and income under a fixed wage rate w_i, per unit of input supplied. The equation of that locus is of course $y_i = w_i r_i^* - w_i v_i$. The second linear locus, TT', which is parallel to Wr^*, represents the result of a transfer WT to individual i, with no change in wage rate. Finally, the concave locus AT' is the result of a combined transfer plus a progressive tax on income which by definition extracts a larger percentage of y_i the higher the level of that income. This locus will, of course, differ from individual to individual, particularly if they do different types of work or differ in skill, or if they supply totally different inputs.

We now can proceed in the usual manner to

*Princeton and New York Universities, and New York University, respectively. Baumol is the author of the body of the paper and Fischer wrote the Appendix. The approach described in this paper has an affinity to a number of recent pathbreaking contributions to tax theory, including work by Peter Diamond and James Mirrlees, A. B. Atkinson, Ray Fair, Martin Feldstein, and Eytan Sheshinski. While the present work is clearly related to theirs, it is perhaps simultaneously more primitive in its approach but more directly germaine to the general issue raised in this paper. For two papers to which this is more directly related, see Bernard Wasow, and Robert Cooter and Elhanan Helpman. We are deeply grateful to Elizabeth Bailey and Alan Blinder for their helpful comments. We must also express our deep gratitude to the National Science Foundation and to the Sloan Foundation for their support of our research.

[1]For example, the disincentive effects of income equalization are recognized throughout the works of the utilitarians of the late nineteenth and early twentieth centuries, who tempered their egalitarian arguments accordingly. Even Karl Marx, who felt that under communism distribution could ultimately proceed in accord with the maxim, "From each according to his ability, to each according to his needs," concedes that in the early stages of communism when it is "... still stamped with the birth marks of the old society from whose womb it emerges" (p. 17), a worker will have to be paid in accord with his contribution: "The same amount of labour which he has given to society in one form he receives back in another" (p. 18).

FIGURE 1. BUDGET LINES, WITH AND WITHOUT TAXES AND TRANSFERS

FIGURE 2. INEQUALITY DESPITE COMMON BUDGET LINE FOR TWO PERSONS

see how such taxes and transfers affect the individual's income, leisure choice, and the degree of inequality of income distribution.[2]

The discussion assumes that, throughout the relevant range, no one is sated in either income or leisure. Consequently, it is assumed that any person's after-tax budget curve between income and retained input is nowhere positively sloping, since with no one sated in either income or retained inputs, all but the highest point on a positively sloping segment will be irrelevant.

PROPOSITION 1: *If different individuals have different budget curves between income and retained input (after transfers and taxes) or if they share a common budget curve which has a negative slope over any interval, then there exist utility functions for which the final distribution of income will be unequal.*

The argument is trivial, and is illustrated in Figure 2 with points I and J the equilibria (with differing income levels y_i and y_j) of two individuals having the corresponding indifference curves. In the diagram they are assumed to face the same budget line in order to show that the problem arises even in this extreme case in which strong equalizing influences are built into the circumstances. A fortiori, the same conclusion *must* hold if their budget curves are different—which is, of course, the major objection of egalitarians to current distribution arrangements.

COROLLARY to Proposition 1: *In the large-numbers case*[3] *only a budget line which is the same for all individuals and is horizontal, that is, one which satisfies $y = k$, can guarantee equality of incomes.*

[2]Note that here we are seeking to approximate the issues considered in reality by phrasing the analysis in terms of equality of (observable) money incomes rather than by equality of subjective utilities.

[3]All the indifference curves here and in what follows will always be defined in terms of an individual's leisure and his or her *direct* after-tax income, assuming that the size of any transfer payment received by an individual is unaffected by the amount of labor he supplies. Any *indirect* effect of an individual's input supply on the size of the pool of total tax revenue and, through this, on the size of the transfer payment he receives, is not included. When the number m of individuals involved is small, this omission may not be justifiable since the $1/m$th of his tax contribution which flows back to him may not be negligible. But when the number of individuals is large, as it will be in any relevant situation in reality, this indirect effect must be negligible, and must be ignored just as we always ignore the effect of the output of an individual supplier upon market price in a competitive industry. We are grateful to an anonymous referee for raising this issue.

PROPOSITION 2: *In an economy with a large number of persons, if individuals are not sated in incomes or retained input, and if zero input supplies yield zero output, then any system of progressive taxes and transfers that guarantees*[4] *equality of incomes must yield zero output in the economy.*

That is, the egalitarian solution will not be feasible if any output is required for survival! The reason, of course, is that in the process of guaranteeing absolute equality, all incentive for supplying any input is removed.

PROOF:

Let r_i^* be the quantity of resource available to individual i, and let v_i be the amount that person retains for self use. Then if (v_i', y) is any point on the income-retained input budget line, with $v_i' < r_i^*$, by the Corollary to Proposition 1, (r_i^*, y) will also lie on that budget line. But by the assumption of nonsatiety we must have, with utility function u^i, $u^i(r_i^*, y) > u^i(v_i', y)$, that is, the individual will always prefer to retain more of his input. Consequently, that individual will always choose to supply the quantity of input $r_i = r_i^* - v_i = r_i^* - r_i^* = 0$. Since this also holds for every other individual in the economy, the economy's total output must be $y = 0$.

This is illustrated in Figure 3 where by the nonsatiety assumption every indifference curve must have a negative slope and higher curves must be preferred to lower ones. Hence, with the horizontal budget line EE' which, by the Corollary to Proposition 1, is the only one that can guarantee equality of incomes under progressive taxation and transfers, the zero input supply point E' must always be optimal. This is the free-rider problem carried to its ultimate extreme.[5]

FIGURE 3. HOW HORIZONTAL BUDGET LINE INDUCES ZERO INPUT SUPPLY

Of course, there may be special situations in which equality can be achieved, using no more than taxes and transfers, without reducing output to zero. An obvious case, already noted, is that in which everyone's input supply function is identical and everyone's input of the same quality. A slightly less restrictive case is that in which input supply functions differ, but there exists a wage rate at which *everyone* supplies the same quantity of (homogeneous) input.

In an economy with many individuals such cases are extremely implausible. We may therefore postulate that *normally* there will be differences in input supply functions and input productivities sufficient to lead to the conclusion of Propositions 1 and 2. That is, these differences will normally be sufficient to imply that only a uniform horizontal budget

[4]Of course, in special situations such a guarantee is unnecessary. If, for example, everyone had equal abilities and identical utility functions, equal incomes would result without any transfer or progressive tax. A bit more will be said about such possibilities presently.

[5]Again note that the free-rider problem (as always) is weaker in the small-numbers case. In a ten-person economy a work agreement may be reached even though payment is independent of productivity. This is, no doubt, part of the explanation of the success of the kibbutzim in Israel where precisely that arrangement holds (see, for example, Haim Barkai). But in the generally more relevant case of large numbers neither social pressure nor a significant marginal payoff to the individual from the contribution of his labor to the size of the total pie will serve to elicit labor from him.

line can produce income equality via taxes and transfers, so that equality can only be achieved at the cost of reducing output to zero.

Let us make this assumption explicit—the premise that indifference curves in retained income-output space (or, equivalently, the income supply functions) of the *m* individuals are such that, whenever the tax schedule (AT' in Figure 1) has a negatively sloping portion, there will be at least two individuals with different incomes. For brevity we will refer to this as the *normal variation property*, meaning that variations in the abilities and utility functions of the members of the group are *normal*.

At this point it becomes useful to describe the tradeoff between output and degree of equality employing a concept we will call the *income distribution frontier*, which is a slight modification of the standard production-possibility locus. Our frontier represents the maximal vectors of *money incomes* which the economy can provide to the individuals who compose it. We find later that the position of this frontier will shift when there is a change in the means used to redistribute incomes.

In the two-person case, which we will use for convenience in diagramming, this graph,[6] which may be written $y_2 = G(y_1)$, shows the maximal income available to individual 2, given the income provided to individual 1. The set of points in (y_1, y_2) space representing outputs that lie on or below and to the left of the boundary, we call the region of feasible incomes.

PROPOSITION 3: *If only progressive taxes and transfers are used to redistribute incomes and the feasible region in y_i space is bounded and the derivatives of its boundary are continuous, and if the normal variations property holds, then a feasible region that contains any positive solution $y_1^* > 0, y_2^* > 0$ for any pair of individuals, 1, 2, must have a frontier that contains some positively sloping segment.*

[6]For an ingenious geometric construction that derives the boundary from two individuals' behavior and preference relations, see Bernard Wasow.

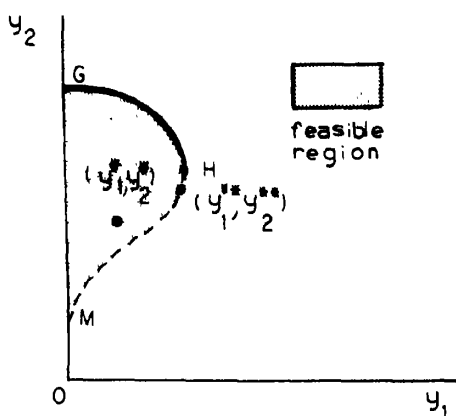



FIGURE 4 OUTPUT DISTRIBUTION FRONTIER

PROOF:

By Proposition 2 (and the normal variations property) the egalitarian solution must involve $y_1 = y_2 = 0$. Moreover, the boundary must then contain a point (y_1^{**}, y_2^{**}) with $y_1^{**} \geq y_1^*$ and $y_2^{**} \geq y_2^*$. Hence, by continuity of derivatives, the segment of the frontier connecting $(y_1, y_2) = (0, 0)$ and $(y_1, y_2) = (y_1^{**}, y_2^{**})$ must contain a positively sloping segment (MH in Figure 4).

The reason for the positively sloping segment is not difficult to explain. With both individuals reduced to zero incomes by a tax arrangement which is absolutely egalitarian, there must in any viable community be an alternative tax arrangement which is at least slightly less egalitarian and which yields a nonzero income to each individual. This means that the move from the first to the second of these arrangements must increase the income of each individual, and hence the segment of the output distribution frontier connecting the two corresponding points will, on our continuity assumption, contain a positively sloping portion ($\Delta y^1 > 0, \Delta y^2 > 0$).

COROLLARY to Proposition 3: *Using only progressive income taxes and transfers to redistribute incomes, the efficient set in y_1, y_2 space must be discontinuous.*

PROOF:

By Proposition 2 the origin is an efficient point. Since no more than one point on a positively sloping segment of the boundary of the feasible region can be *efficient*, the result follows.

Thus, in Figure 4, while *GHMO* is the income distribution frontier, the efficient locus is composed of the segment *GH* together with the origin.

It should be noted that usually, starting from an initial point at which, say, $y_2 > y_1$, no normal system of income taxes and transfers will take us to a point at which $y_1 > y_2$. It follows that the feasible region will usually contain points on only one side of the line $y_1 = y_2$, as shown in Figure 4. That is, there is no way in which progressive taxes and transfers alone can produce an interchange in the position of the rich and the poor à la the Prince and the Pauper.[7] The reason is that any taxes capable of producing such an interchange must produce disincentives to labor even stronger than those necessary to produce perfect equality and so they too must disrupt the economy's productive process.

II. Digression: Goals Alternative to Perfect Equality

We can use our construction to examine the implications of some objectives other than perfect income equality under a regime of progressive taxes and transfers. We will consider only three such goals, though others can easily be suggested:[8]

1) Maximization of social output $y = \Sigma y_i$.

2) Maximization of a social welfare function[9] $U(y_1, \dots, y_m)$ (an "antipoverty" welfare function) which gives weight to equalization of incomes by placing a greater value upon an increase in the income of a poor man. Thus, writing $U_i = \partial U/\partial y_i$, etc.,

$$(1) \qquad U_i = U_j \text{ for } y_i = y_j \qquad i, j = 1, \dots, m$$

and

$$(2) \qquad U_{ii} < 0 \qquad i = 1, \dots, m$$

3) John Rawls' maximal-justice solution, which requires maximization of benefit to the individual who is in the most disadvantageous position.

The three solutions are readily depicted in Figures 5–7. Figure 5 shows point *M* corresponding to *y*, the highest of the attainable iso-income loci $y_1 + y_2 = k$. This figure also depicts the 45° ray corresponding to absolutely equal distribution of incomes, showing that the only feasible point on this ray is the origin. Figure 6 depicts the point *A* that maximizes the welfare function[10] $U(\cdot)$, whose indifference curves I, I', and I'' are symmetric about the 45° ray because of the premise (2) that for equal incomes, $U_1 = U_2$. We would expect *A* to lie closer to the 45° line than *M* does. For the output-maximization solution (point *M*) requires (taking, for example, $y_2 > y_1$ as in the case in the diagrams) $dy_2/dy_1 = -1$, while the welfare-maximization solution requires $dy_2/dy_1 = -U_1/U_2$ where, by (1) and (2), $|U_1/U_2| > 1$, since $y_2 > y_1$, so that *if the efficient locus is concave*, *A* must lie closer than *M* to the vertical point *V*.

Finally, Figure 7, following Phelps, deals

[7]For a beautiful analysis of Rawlsian solution using the utility distribution frontier depicting it as lying entirely on one side of the 45° ray, see Edmund Phelps, sec. I. For a simulation in utility space that contains such a frontier, showing it going through the origin at the 45° ray, see Cooter and Helpman, Figure I, p 660.

[8]Cooter and Helpman distinguish several other such criteria and show for each of them how the optimal position on the distribution frontier can be calculated. They include an elitist solution (maximal income to the most productive individual), the Benthamite solution, which maximizes the sum of utilities, the Nash solution, which maximizes their product, and the "democratic solution," which maximizes the income of the class of median ability. They also define an egalitarian solution that minimizes the Gini coefficient (see their p. 658).

[9]For sophisticated analyses of such optimal solutions see Mirrlees and Peter Wagstaff.

[10]The indifference curves I, I', and I'' of our social welfare function must have the curvature usually assumed of indifference curves. This follows by assumption (2) that the second partials U_{ii} of the welfare function are negative, if we also take the cross partials U_{ij}, $i \neq j$, all to be zero since the y_i for any individual i is independent of the earnings of any other individual. For along any indifference curve, where $U(\cdot) = k$, we have $dU = U_1dy_1 + U_2dy_2 = 0$, so that $dy_2/dy_1 = -U_1/U_2$. With $U_{21} = 0$, we then obtain $d^2y_2/dy_1^2 = -U_{11}/U_2 > 0$, giving the required curvature of the indifference curve.

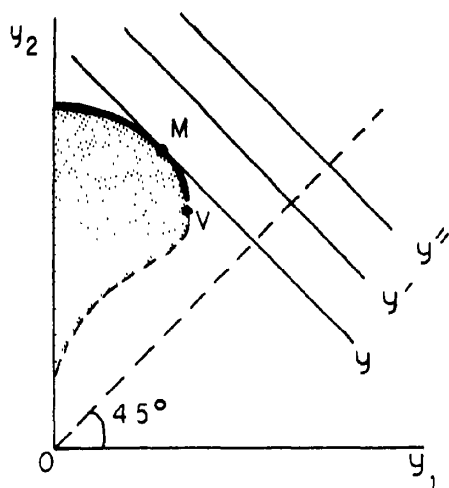



FIGURE 5. INCOME MAXIMIZATION

with the Rawlsian case. Rawls' criterion can be taken roughly to imply that a change is desirable if and only if it increases the income of the most impoverished individual in the community. This yields a social welfare function whose indifference curves are essentially replications of the axes with the origin shifted outward along the 45° ray. For this tells us that starting from a point of equality such as E, a gain to either party *alone* is not an improvement. However, from any point (such as V) to the left of the 45° ray, where individual 1 has the lower income, any

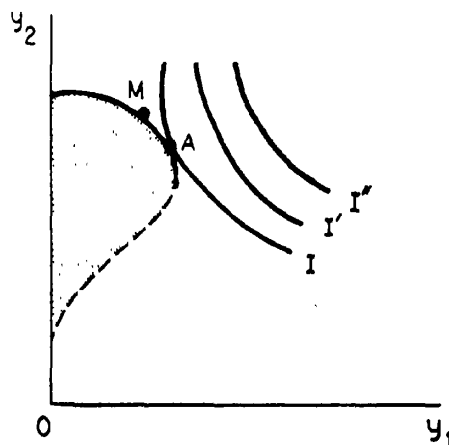



FIGURE 6. ANTIPOVERTY WELFARE MAXIMIZATION

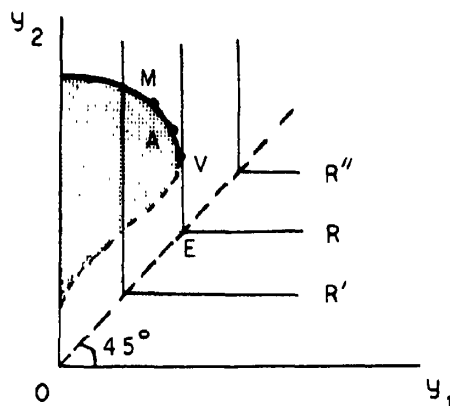



FIGURE 7. RAWLSIAN SOLUTION

increase in 1's income alone gets us to a higher indifference curve.

Here we see that the Rawlsian solution will be given by the vertical point V on the boundary of the feasible set. It will be the most egalitarian point on the upper segment of the efficient locus, but from Proposition 2 it follows that it will *not* be perfectly egalitarian when distribution is achieved by transfers and progressive taxes.

III. An Alternative Approach to Income Equalization

The main and most surprising result in our discussion so far is that with the methods of redistribution currently used, complete equalization of income would, at least in theory, bring the entire productive mechanism to a halt. One is led to surmise on a continuity assumption from this extreme case that very large output losses are also likely to accompany any attempt to get anywhere *very close* to equality.

All this is not meant to imply that equality of incomes, or some close approximation to it, is an impossible goal. Rather, it means that the methods we are currently using for the purpose, namely a combination of the progressive income tax and transfers, may simply not be up to the task, though it may be the best way to institute a more modest redistribution. We will now describe a theo-

retical solution which, while it can only be approximated in practice, in pure form it does make an egalitarian solution possible without exacting a high cost in terms of foregone income. The main purpose of this abstract concept is perhaps to be taken to be akin to that of an existence theorem—that is, it is intended to show that we are not necessarily wasting our time in looking for new and more heterodox means to reduce inequality without imposing an unacceptably high income loss upon society.[11]

But, as will be discussed later, there are ways in which this redistribution procedure can be approximated in practice. To describe our alternative instrument for redistribution, let us set the policymaker free to determine a fixed wage rate w_i per unit of input supplied by an individual i, where w_i may differ from the wage rate w_j for any other individual j in any way the policymaker considers appropriate. In other words, the policymaker is given complete liberty to set *discriminatory input prices*, supplier by supplier, provided they satisfy the (inequality form of the) feasibility condition requiring total output to equal or exceed the sum of the resulting incomes.

Under such an arrangement the policymaker can achieve an egalitarian solution as follows. In Figure 8 let $\bar{y}$ be an income level which the policymaker hopes, tentatively, to achieve for every earner. By trying alternative budget lines Wr_i^*, $W'r_i^*$, etc., with neither transfers nor progressive taxes, he generates an offer curve RR' for individual i from the points of tangency between the budget lines and i's indifference curves. The point E at which the offer curve crosses the horizontal line through target income level $\bar{y}$ gives us the required discriminatory wage rate w_i, which is the absolute slope of budget line $W'r_i^*$ through point E. Then, if the corresponding

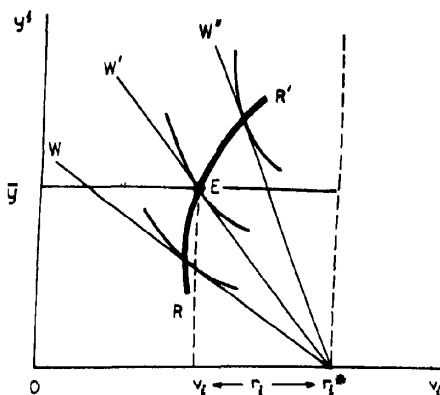



FIGURE 8. EGALITARIAN SOLUTION UNDER WAGE DISCRIMINATION

input supplies satisfy the feasibility condition requiring total output to equal total income (see equation (3) below), then this is the final solution. Otherwise, the policymaker must repeat the process trying another value of $\bar{y}$ until a feasible solution is found.

To describe the process algebraically we use the following notation:

$r_i = g^i(w_i)$ = individual i's input supply function

$y = f(r_1, \ldots, r_m)$ = the production function

y_i = individual i's income $= w_i g^i(w_i)$

We deal, for simplicity, with a single product economy in which each of the persons supplies a single input (which may vary from individual to individual so that there can be $n \leq m$ different inputs). The quantity of input supply by any one individual i varies only with the price w_i he receives per unit of input.

Algebraically, this egalitarian solution is then found by solving the condition for the feasibility of the proposed egalitarian wage $\bar{y}$,

$$(3) \qquad m\bar{y} = \sum y_i = f(r_1, \ldots, r_m)$$

simultaneously with the $m - 1$ conditions

$$(4) \qquad y_i = y_m$$

that is, $w_i g^i(w_i) = w_m g^m(w_m)$

$$i = 1, \ldots, m-1$$

after rewriting (3) as

$$(3') \qquad \sum_{i=1}^{m} w_i g^i(w_i) = f[g^1(w_1), \ldots, g^m(w_m)]$$

[11] As pointed out by a referee, we all know that this can be done, in theory via lump sum taxes, though it is hard to imagine how a tax can guarantee equality and yet be lump sum. That is, if each person is guaranteed the same income as everyone else's (even if only in terms of expected value) we run into the problem of output loss indicated by Proposition 2. With such disincentive effects, the redistribution payments by definition cannot be lump sum.

Note that this solution yields perfect equality without tempting anyone to reduce his input supply to zero. As Figure 8 shows, under the discriminatory price arrangement no one is *guaranteed* the income level $\bar{y}$. A person receives that income level at his assigned wage rate *only* by providing the amount of input indicated by his supply function. If, instead, he chooses to supply a zero input (point r_i^*) he receives a zero income. It is true that equality requires $w_i = k/r_i$, that is, it requires wage rates to vary inversely with i's input supply. But our redistribution process does not proceed by adjusting i's wage rate upward each time his input supply decreases. Rather, it presumes that i's input supply function $g^i(w_i)$ is known to the policymaker and that he uses it to solve for that wage rate at which (4) will be satisfied by i's known and predictable wage response, that is, the policymaker solves the set of m simultaneous equations given by (4) and (3′), obtaining the m wage rates $w_1^e, \ldots, w_m^e$. The members of the economy are then left to themselves to adjust to these wage rates, which we are assured by conditions (3′) and (4) will *automatically* yield the equal income distribution that is desired.

Thus, unlike the case in which transfers and progressive taxes are used to redistribute incomes, the discriminatory wage arrangement need not yield a possibility locus that goes through the origin, or even near it, even if the variations in input supplies and abilities are "normal," as is proved in the Appendix.[12] The frontier may even be concave throughout, as one might have expected on first examining the subject.[13]

[12]Of course, the discriminatory wage arrangement must affect the prices of final goods and the allocation of resources, though we have no idea of the likely magnitude and direction of these effects. The shape and position of the possibility locus will therefore also be affected.

[13]A comparison of Figures 9 and 10 in the Appendix indicates what happens under a policy of discriminatory wage rates to the solutions examined in Figures 5–7. The Rawlsian solution (V) is now identical with the egalitarian solution (E). The antipoverty solution (A) will lie between E and the output-maximizing solution (M), since under A a relatively large gain to one individual with only a small loss to the other will be preferred to complete equality at the expense of a lower total output. But output under A will still be lower than under M, because the social welfare function will generally not consider the last extractable iota of gain in output to be worth the required increase in inequality.

IV. The Role of the Utility of Leisure

The concreteness of the analysis as described up to this point has been achieved at the cost of what can be interpreted as an illegitimate oversimplification. For though we have claimed to be dealing with a single output economy, in fact a second output, leisure (input retained by suppliers for personal use), has unavoidably entered the picture. After all, unattractive wage rates decrease input supplies only because their suppliers obtain some utility by keeping additional amounts for themselves. The tradeoff between increased equality and output may therefore, in principle, be no tradeoff at all. For reduced output means increased leisure, and it is at least conceivable that as a result increased leisure may entail no loss in the welfare of any individual even though it reduced some people's or even everyone's flow of material goods.[14]

How do we know that under progressive taxation complete equality is likely to entail a loss of *utility* even if it substitutes leisure for material output? The answer lies in Proposition 2 which shows that with the indifference maps postulated, if tax arrangements are *absolutely* egalitarian in terms of purchasing power, society will end up with zero employment of resources and zero material goods output. We can invoke the indifference maps to show that this is surely no optimal state of affairs, but such an exercise hardly seems necessary.[15]

[14]The analysis of the possibility frontiers is easily reformulated to take this issue into account. While there is no way we can add together output and leisure we can deal with them both with the help of an ordinal utility possibility locus. For this we must take our axes to represent, rather than y_1 and y_2, the two ordinal utilities, u^1 and u^2. As usual, we have no way of comparing these two magnitudes, which can each legitimately be modified by any monotone transformation.

[15]Thus, it follows that at least part of the utility frontier under progressive taxes must lie inside the frontier corresponding to wage discrimination. The use of a utility frontier rather than an output frontier does, however, inhibit our analysis to some degree. Taking redistribution, as a matter of realism, to mean redistribu-

V. Conclusion: Toward Practical Policy

Our theory has already told us several things pertinent for policy. First, it has shown that however helpful progressive taxation may be in achieving *small* increases in equality, it can be inferred (via a continuity assumption) from our analysis of the extreme case of perfect equality,[16] that it can impose a terrible burden upon the productive mechanism of any economy where it is used to eliminate all but minor differences in incomes. The implications for countries that have already gone far along this path may be worth noting.

Second, we have seen that, at least in theory, there may be a way out of the problem. The perfectly discriminatory prices of inputs that our analysis used to illustrate such alternative procedures cannot, of course, be put into practice in all their detail. One cannot vary wage rates, input supplier by input supplier, in accord with fine-tuned calculations based on the individual supply functions. However, some ingenuity may be able to produce viable approximations to the discriminatory wage policy which we have shown to be capable of getting us closer to equality without necessarily causing a catastrophic loss in output.

As with any arrangement involving effective price discrimination, some knowledge of relative elasticities is required, and it must be difficult to shift goods or services from lower- to higher-priced markets. Both of these issues can perhaps be dealt with by discriminating not among individuals but among broad groups of income earners. If doctors and ditch diggers are fairly noncompeting groups, with little movement of labor from one to the other, a national wages policy which discriminates between them may be entirely feasible. Moreover, for such broad groups the estimation of labor supply functions is surely not out of the question. One can supplement wage rates for the one and limit wages for the other without expecting a large attrition in the supply of doctors by migration into ditch digging. There are obvious problems besetting this process and any brief discussion of such a complicated matter is inevitably condemned to naiveté. The approach will run into trouble if large international differences in doctors' wages persist, since doctors can migrate abroad. High ditch digger wages will discourage demand for the unskilled unless the wages are supported by public subsidy. But the fact is that interoccupational wage ratios have been changed in a number of countries through government policy—national health programs providing a prime example. Thus, it *can* be done. We have not argued that it should be done, but have suggested that interoccupational wage differences can be and have been influenced in practice, and that these represent one measure that is in principle capable of producing substantial decreases in inequality without nearly the disincentive effects of means we have traditionally used for the purpose.

All economists are aware that egalitarian measures are likely to exact a cost in terms of a loss in national income. What seems not to have been considered before is the possibility that heterodox means may be able to limit this cost considerably even if the desired increase in equality is substantial.

Appendix: An Example Proving the Discriminatory Frontier Need Not Go Through the Origin

This section will illustrate by means of a concrete example the income vs. equity tradeoffs that are possible under the two redistribu-

tion of output (purchasing) power, not equalization of utility (who ever heard of a progressive tax upon psychic utility, and what would it mean?), the solution values corresponding to our four prototype solutions—the output maximum, the welfare maximum, the egalitarian and the Rawlsian solutions—remain the same as before. However, in utility space they all lose their simple representations. Since in that space the utility indices for the two individuals are not comparable, the 45° line loses its meaning as a standard of equality, and so our devices for the representation of the egalitarian and the Rawlsian solutions no longer work. Similarly, a negatively sloping 45° line no longer represents either a fixed total output or a fixed total utility for the two parties. This is a problem which seems to have been overlooked in previous graphic analyses of the subject.

[16]This is a Marshallian view of matters: *Natura non facit saltum* (or in the words of a fortune cooky "Nature does not proceed by leaps").

tion measures we have discussed: progressive taxation and differentiated wage rates. By its example, it will prove[17]

PROPOSITION 4: *Under discriminatory wages it is not impossible to achieve absolute equality without zero outputs, even for normal input supply curves. Indeed, at the point of equality the boundary of the feasible set in output distribution space can be negatively sloping and concave.*

Assume an input supply function for each of two individuals which, as the wage rate rises, increases rapidly at first, then more slowly and smoothly approaches a maximum. Simple functions with this property are given by[18]

$$(5)\quad g^i(w_i) = \begin{cases} r_i^*(1-(1-w_i)^2) & \text{if } 0 \le w_i < 1 \\ r_i^* & \text{if } w_i \ge 1 \end{cases}$$

with $r_1^* = 1$ and $r_2^* = 2$. The input supply functions (5) are normal in the sense that no scheme of taxation can yield equal income without reducing output to zero.[19]

Let us now determine what total output y will be produced by the two individuals under various schemes of taxation and redistribution. Consider the following simple taxation scheme: Both individuals pay a tax at the same rate t, and the revenue obtained in this way is distributed equally between the two. (Both, of course, are paid at the same wage rate $w = w_1 = w_2$ under this first set of policy measures.) If $t = 0$, no redistribution will take place. If $t = 1$, the entire output is taxed away and distributed equally. As was shown in Proposition 2, such a policy results in a maximal disincentive effect and yields a zero output. Let us now analyze intermediate cases.

Consider the simple production function[20]

$$(6)\qquad y = f(r_1, r_2) = r_1 + r_2$$

which permits a maximum wage rate of $w = 1$. If a tax rate t is applied, then the after-tax wage rate is $w = 1 - t$ and from (5) it follows that the tax revenue, T, which is to be distributed equally, is given by

$$T = r_1 + r_2 - w(r_1 + r_2) = 3t(1 - t^2)$$

This yields, as the incomes obtained after taxes and transfers,

$$(7)\quad y_1 = r_1(1-t) + T/2 = (1-t^2)(1+t/2)$$
$$y_2 = r_2(1-t) + T/2 = (1-t^2)(2-t/2)$$

for $0 \le t \le 1$.

As the tax rate t is gradually increased from 0 to 1, the income y_1 and y_2 move from

[17] As Schlomo Maital once reminded Baumol, there is a Yiddish proverb that states, "for example is not a proof." However, here we obviously have an exception, since Proposition 4 merely claims that its result is logically possible. Indeed, the example goes further, showing that one needs no implausible relationships to yield the result asserted.

[18] It can be shown that a series of indifference curves in income-leisure space for each individual which implies the assumed input supply behavior is given by

$$u^i(y_i, v_i) = y_i + v_i - \frac{2}{3} r_i^* \left(\frac{v_i}{r_i^*}\right)^{3/2} = \text{constant}$$

Transfer payments do not affect the individuals' input supply in this example.

[19] To see this, we note that any tax and transfer system in which the transfer payments depend on income is always equivalent to another scheme under which everybody receives the same transfer payment and only the tax levied depends on income. Now, for income equality, both individuals would have to receive the same after-tax wage income. Let $w = w_1 = w_2$ be the wage rate, $t(wr)$ be the progressive income-dependent tax rate, and $h(r) = wr[1 - t(wr)]$ the after-tax wage income. We have assumed that $dh(r)/dr > 0$ for $0 \le r \le \max(r_1^*, r_2^*)$, i.e., there is never any penalty for additional input supplied (see Figure 1 and the discussion preceding Proposition 1). Thus, to obtain equal incomes the two individuals would have to supply the same amount of input. Since both face the same after-tax wage income schedule $h(r)$, they would also have to exhibit the same preference for leisure at that point. But from $du^i/dr_i = -du_i/dv_i = [(r_i^* - r_i)/r_i^*]^{1/2} - 1$ and $du^i/dy_i = 1$ we find that since $r_1^* \neq r_2^*$ we always have $dy_1/dr_1 \neq dy_2/dr_2$ for $r_1 = r_2$, except when $r_1 = r_2 = 0$; i.e., the two individuals never have the same marginal rate of substitution between input supply and leisure except when output is zero. Note that the effect of the input supply on the size of the transfer payment is not considered here. (Compare the fn. to the Corollary to Proposition 1.)

[20] This production function would apply, for example, in a case where the input is labor and the output is measured in terms of its labor content.

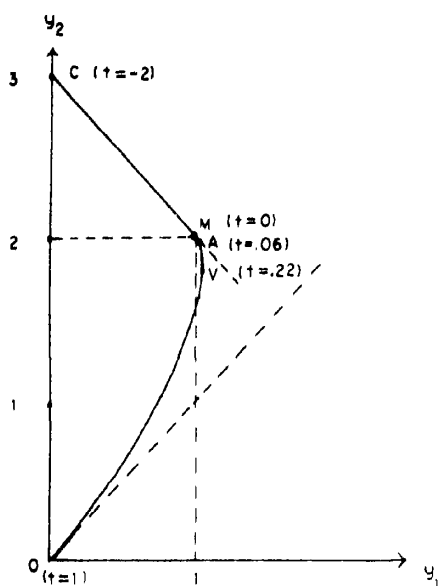



FIGURE 9. THE OUTPUT DISTRIBUTION FRONTIER IN y_1, y_2 SPACE, UNDER VARIOUS DEGREES OF TAXATION AND REDISTRIBUTION

FIGURE 10. THE OUTPUT DISTRIBUTION FRONTIER IN y_1, y_2 SPACE, UNDER VARIOUS FEASIBLE COMBINATIONS OF DIFFERENTIATED WAGE RATES

point M in Figure 9 through A and V to the origin.[21] The maximum output (point M) is reached when the tax rate is $t = 0$.[22] A welfare-maximizing solution (point A) was determined utilizing as the welfare function the product $y_1 \cdot y_2$. This point is reached

[21]With a negative tax rate, the wage rate would be increased to more than 1, and a lump sum would have to be extracted from each individual to compensate for the resulting deficit. The income-distribution frontier would then be extended from point M to C in Figure 9, but this hardly seems realistic.

[22]Since there are no externalities in this model, and there is no advantage in mutual cooperation, it is natural to find that a pure *laissez-faire* strategy with equal wage rates and no taxation yields the output-maximizing solution

TABLE 1—EFFICIENT INCOME COMBINATIONS UNDER TWO TYPES OF INCOME REDISTRIBUTION POLICIES

	Output Maximization (M)	Welfare-Maximization Solution (A)	Rawlsian Solution (V)	Income Equalization (E)	Maximization of Individual 1's Income (D)	Maximization of Individual 2's Income (C)
Taxation with Redistribution						
Tax Rate	0	.06	.22	1	.22	−2
Transfer Payment	0	.09	.31	0	.31	−3
y_1	1	1.03	1.06	0	1.06	0
y_2	2	1.96	1.81	0	1.81	3
$y = y_1 + y_2$	3	2.99	2.87	0	2.87	3
Differential Wage Rates						
w_1	1	1.33	1.45	1.45	1.77	.42
w_2	1	.83	.77	.77	.42	1.19
y_1	1	1.33	1.45	1.45	1.77	.28
y_2	2	1.62	1.45	1.45	.56	2.38
$y = y_1 + y_2$	3	2.95	2.90	2.90	2.33	2.66

when $(1/y_1)(dy_1/dt) = -(1/y_2)(dy_2/dt)$. The various solutions are summarized and compared in Table 1.

Let us now explore what the efficient income set looks like for this example if differentiated wage rates, rather than taxation and redistribution, are used as policy instruments. If individual i is paid at a wage rate $w_i < 1$, she produces more than she earns and leaves a surplus of

$$S_i = r_i - w_i r_i = (1 - w_i) r_i^* (1 - (1 - w_i)^2)$$

If $w_j > 1$, she has a deficit

$$S_j = (1 - w_j) r_j = -(w_j - 1) r_j^* < 0$$

This deficit must then be compensated for by the surplus of the other individual. For $0 \leq w_2 \leq 1$ we obtain

$$(8)\quad y_1 = w_1 r_1 = w_1 \cdot 1 = 1 + 2(1 - w_2)(1 - (1 - w_2)^2)$$

$$y_2 = w_2 r_2 = 2w_2(1 - (1 - w_2)^2)$$

A similar pair of formulas applies when $0 \leq w_1 \leq 1$. The resulting output distribution frontier is shown in Figure 10. Table 1 gives a summary of the various solutions obtained.

A comparison of the income combinations that are feasible under taxation with redistribution and under wage rate differentiation confirms that under rather natural assumptions about individuals' behavior, the second set of policy measures can permit far more equitable income distributions without reducing total output substantially. Indeed, in our example, complete equality merely reduces total output to 2.9 from its maximal value of 3.0.

REFERENCES

A. B. Atkinson, "How Progressive Should Income Tax Be?," in Michael Parkin, ed., *Essays on Modern Economics*, London 1973.

H. Barkai, "Productivity in the Kibbutz—Small Scale Enterprise Outside the Profit Sector," in William Baumol, ed., *Public and Private Enterprise in a Mixed Economy*, London forthcoming.

R. Cooter and E. Helpman, "Optimal Income Taxes for Transfer Payments," *Quart. J. Econ.*, Nov. 1974, *88*, 656–70.

P. A. Diamond and J. A. Mirrlees, "Optimal Taxation and Public Production: II," *Amer. Econ. Rev.*, June 1971, *61*, 261–78.

R. C. Fair, "The Optimal Distribution of Income," *Quart. J. Econ.*, Nov. 1971, *85*, 551–79.

M. Feldstein, "On the Optimal Progressivity of the Income Tax," *J. Publ. Econ.*, Nov. 1973, *2*, 357–76.

Karl Marx, *Critique of the Gotha Program*, Moscow 1937.

J. A. Mirrlees, "An Exploration in the Theory of Optimum Income Taxation," *Rev. Econ. Stud.*, Apr. 1971, *38*, 175–208.

E. S. Phelps, "Taxation of Wage Income for Economic Justice," *Quart. J. Econ.*, Aug. 1973, *87*, 331–54.

E. Sheshinski, "Income Taxation and Capital Accumulation," *Quart. J. Econ.*, forthcoming.

P. Wagstaff, "A Benthamite Wages Policy," *Rev. Econ. Stud.*, Oct. 1975, *42*, 571–80.

B. Wasow, "Optimal Income Distribution with Variable Factor Supplies: A Graphical Exposition," unpublished paper, Center Appl. Econ., New York Univ. 1975.

On Fiscal and Monetary Policies and the Government Budget Restraint

By CARL F. CHRIST*

Recognition of the government budget restraint has altered the way economists think about macroeconomics. This is especially true concerning what policy options are available, stability conditions, long-run multipliers, and the nature of equilibrium. One purpose of this paper is to describe and illustrate the way the government budget restraint affects macro-economic analysis.

Several writers have found that a particular monetarist type of policy is or may be unstable, to wit: fix government expenditures and tax-transfer schedules, make the money stock grow at a constant moderate rate, and issue or retire government bonds to match any budget deficit or surplus. A second purpose of this paper is to show how the stability of such a policy can at least be made possible for some parameter values, by including government interest (gross of income tax) in the government expenditure variable that is fixed.

I. The Role of the Government Budget Restraint

In this section, before embarking upon mathematical seas, I describe the distinguishing features of macro-economic analysis in the presence of the government budget restraint (*GBR*). The *GBR* is the requirement that the total of government expenditure for all purposes must equal the total of financing from all sources, including printing money, thus:

We consider the *GBR* of the federal government, excluding state and local governments because they cannot print money (they are accordingly included in the private sector). We consolidate the Federal Reserve with the federal government because the effects of their separate and joint policies depend on the actions of their consolidated sector vis-à-vis the rest of the economy, independent of any additional transactions that they may undertake between themselves alone. We consider only a closed economy, ignoring gold and foreign exchange reserves, because of space limitations. We consolidate the banking and the private nonbank sectors for simplicity.

The most fundamental implication of the *GBR* is that the authorities cannot fix arbitrary paths for all of the macro-economic policy variables at once. At least one policy variable must have its path endogenously determined by the joint action of the *GBR* and the economy's structure. For instance, if the authorities fix constant-level paths for purchases of goods and services, tax rates, transfer rates, and privately held government debt, then the path of the high-powered money stock will be determined by the *GBR* and the economy: Whenever income and the price level are such that the budget is in deficit, high-powered money must be issued to finance it; similarly, when the budget is in surplus or is balanced, the high-powered money stock must decline or be constant.

*Johns Hopkins University. I am glad to acknowledge helpful discussions on the subject of this paper with Willem H. Buiter, Frank De Leeuw, Chi-Hung Leung, Louis Maccini, Bennett McCallum, Jürg Niehans, William Poole, Jerome Stein, and Naoyuki Yoshino. The research underlying the paper was supported in part by an *NSF* grant to the Johns Hopkins University. The views expressed here are not necessarily those of *NSF*. Any errors of course are my own.

The authorities can decide which one of the N policy variables is to be endogenous, and can then fix paths for the other $N-1$. Or they can permit two policy variables to vary endogenously by fixing the ratio between them (for example, the ratio of high-powered money to privately held government debt), and fix paths for the other $N-2$.

It is not possible to change just one policy variable from its previous path, leaving all others on their previous paths. If one is changed, at least one other must be changed to satisfy the *GBR*. For example, if currency is dropped from airplanes during a short interval, not only does the high-powered money stock increase, but transfers become temporarily large during that interval.

The path of the economy after a policy change depends not only on which two (or more) policy variables are changed initially, but also on the choice of which policy variable is to be endogenous afterwards. For example, the effect of an open market purchase of bonds, if private bondholdings are held fixed afterward and high-powered money adjusts endogenously to cover any budget imbalance, will not be the same as if high-powered money is held fixed afterward and bonds adjust endogenously. The effect will be different again, if *both* bonds and high powered money are held fixed afterwards, and some other policy variable (a government expenditure or tax variable) is endogenous in the subsequent adjustment period.

For simplicity, I will usually confine attention to those policy changes where only two policy variables are changed initially, and where just one of those two is allowed to be endogenous afterward. Such a policy change can be described by specifying one policy variable whose path is to be changed exogenously, and one other policy variable that is to adjust endogenously both initially and afterward.

The questions that can be answered by theoretical macro-economic analysis using the *GBR* are familiar, though the answers are sometimes different when the *GBR* is brought into the picture. Does a particular policy permit the existence of long-run equilibrium, either a static equilibrium position or a dynamic equilibrium path? If so, is it stable? And what are the long-run equilibrium multiplier effects of exogenous changes in each policy variable? The stability conditions and the long-run equilibrium multipliers turn out to depend upon the choice of which policy variable is to be endogenous.

We do not have generally accepted definitions of pure monetary policy and pure fiscal policy. In my view, the quest for such definitions is futile because a change in the money stock or in government purchases or tax transfer rates alone does not describe a policy change adequately: at least two policy variables must be changed initially if policy is to be changed, and then at least one policy variable must be allowed to adjust endogenously afterward. Discussions of macroeconomic policy would be much clearer if each policy change were completely described by specifying which two (or more) policy variables are changed initially, and which one (or more) is chosen to adjust endogenously afterward.

II. A Simple *GBR* Model

To analyze theoretically the implications of the *GBR*, one must specify a complete macroeconomic model. For illustrative purposes, consider the model that is analyzed in my 1978 paper. It is simple enough to understand easily, yet complex enough to permit the study of constant-inflation dynamic equilibrium paths as well as constant-price-level static equilibrium positions. It consists of the following six equations. The meaning of each is described briefly below. A list of notation is provided for reference.

$$(1) \quad y = x + g$$

$$(2) \quad t = V/P - B/P + uB/P + uy$$

$$(3) \quad x = \alpha_0 + \alpha_1 y - \alpha_2 t + \alpha_3/r + \alpha_4 H/P + \alpha_5 B/P + \alpha_6 \pi$$

$$(4) \quad 1/r = \lambda_0 - \lambda_1 y + \lambda_2 H/P - \lambda_3 B/P$$

$$(5) \quad g - t = DH/P + DB/rP$$

$$(6) \quad DP/P = (y/f - 1)\rho + \pi$$

List of Notation

- A_i – characteristic matrix of dynamic system when i is the endogenous policy variable
- B – nominal bond interest, \$/yr, > 0
- D – d/dt, the time-derivative operator
- f – real capacity output, \$/yr, > 0
- g – real government purchases, \$/yr, > 0
- g' – $g + (1 - u)B/P$, \$/yr, > 0
- g'' – $g + B/P$, \$/yr, > 0
- H – nominal high-powered money, \$, > 0
- P – price level, > 0
- r – nominal interest rate, $(yr)^{-1}$, > 0
- t – real taxes less all transfers, \$/yr
- u – marginal tax rate, dimensionless, $0 \leq u < 1$
- V – nominal tax-transfer intercept, \$/yr, < 0
- x – real private expenditure, \$/yr, > 0
- y – real income, \$/yr, > 0
- α_i – private expenditure parameters, > 0 (except possibly α_0)
- Δ – denominator of ϕ in equation (7), > 0
- Δ_i – determinant of equilibrium system when i is the endogenous policy variable
- θ – equilibrium growth rate of B, H, V, P
- λ_i – money demand parameters, > 0 (except possibly λ_0)
- π – expected inflation rate, $(yr)^{-1}$, $= \theta$
- ρ – speed of adjustment of inflation rate, > 0
- ϕ – aggregate demand function in equation (7)
- ϕ_i – partial derivative of ϕ

Equation (1) is the national income identity. It says that real national income y is the sum of private expenditure x (for consumption and investment) plus government purchases of goods and services g.

Equation (2) is the tax-transfer equation. Assume that the macro-economic effects of transfer payments are the same as those of taxes except for sign. Thus we use one equation for real tax receipts less real transfers, all denoted by t. Consider the four terms on the right-hand side in turn: V is nominal autonomous taxes, less autonomous transfers other than government interest. Its real value is V/P. The term B/P is the real value of interest income received by private holders of government bonds. Assume that all bonds are perpetuities paying \$1 a year each in nominal terms; then B stands for both the number of government bonds in private hands, and the nominal interest income thereon. The term uB/P is the induced real tax on government debt interest, at the constant marginal rate u. The term uy is the real induced tax less induced transfers based on real national income, at the marginal rate u. Let us assume that $0 \leq u < 1$ and that V is negative so that the tax transfer system is progressive.

Equation (3) is the real private expenditure equation. All the slope coefficients α_i are positive. Real private expenditure depends positively on real income y, negatively on real taxes less transfers t, negatively on the nominal interest rate r, positively on the real high-powered money stock H/P, positively on real private holdings of government debt B/rP (but note that the equation has been linearized in $1/r$ and B/P for convenience—this is consistent with the linearization that will be performed later for stability analysis), and positively on the expected inflation rate π. Assume that $\alpha_2 \leq \alpha_1 < 1$; the first of these inequalities permits a zero or positive marginal propensity to invest, $\alpha_1 - \alpha_2$.

Equation (4) is the demand equation for real high-powered money H/P. It is obtained thus: First express H/P in the usual way as depending positively on real income y and real private holdings of government debt B/rP, and negatively on the nominal interest rate r; then linearize in $1/r$ and B/P; and then solve for $1/r$. This simplifies the form of equation (7) below. All the slope coefficients λ_i are positive.

Equation (5) is the government budget restraint, stating that the real government deficit $g - t$ must be financed by some combination of printing high-powered money DH/P and borrowing from the private sector DB/rP.

Equation (6) is the inflation adjustment equation. It says that the actual inflation rate DP/P will be equal to the exogenous expected rate π, unless real income is above or below

the full-employment level f, in which case actual inflation will, respectively, accelerate or slow down ($\rho > 0$) as compared with the expected rate π.

This model has six endogenous variables: y, t, x, r, P, and *one* of the five policy variables (B, g, H, u, V). The exogenous variables are f, π, and *four* of the five policy variables. The capital stock is assumed to change so little that it and f can be taken as fixed.

The model generates a static long-run equilibrium position at the full-employment real income level $y = f$, if all exogenous variables are assigned constant values and if zero inflation is expected. It generates a static *real* equilibrium position at $y = f$ and a steady-growth dynamic *nominal* equilibrium path, if all *real* exogenous variables are assigned constant values and all *nominal* exogenous variables (which must be chosen from the set of B, H, and V) are assigned steady-growth paths all with the same growth rate, and if furthermore the expected inflation rate is equal to this common growth rate of the nominal exogenous policy variables. This latter assumption makes long-run sense because in reality expectations about inflation will not long remain different from the actual inflation rate if the latter remains constant. The short-run meaning of the assumption amounts to a rational expectations postulate, for it says that inflationary expectations promptly mirror the decision of the authorities regarding the growth rate of the nominal exogenous policy variables.

Note that the equations are of two types: (5) and (6) are dynamic, each describing how some variable changes through time, either the price level or a stock of assets; (1)–(4) are static, describing short-run equilibrium conditions for given values of the price level and the asset stock whose evolution is prescribed by the dynamic equations.

III. Steps in Analyzing a *GBR* Model

A useful method of analysis consists of the following steps (a)–(g). It is a standard dynamic analysis method, adapted for use with the *GBR*.

(a) Solve the static equations for the short-run equilibrium values of all endogenous variables except those whose time derivatives appear in the dynamic equations, taking the latter and the exogenous variables as given. In a simple *IS–LM* model, this amounts to finding the intersection of the *IS* and *LM* curves. In the foregoing model, (4) is the *LM* curve, and the *IS* curve is obtained from (1)–(3). If the interest rate is eliminated between the *IS* and *LM* curves, the result is the usual aggregate demand curve relating real output negatively to the price level, thus:

$$(7)\quad y = \phi(P, \pi, B, g, H, u, V) = [\alpha_0 + \alpha_3\lambda_0 + \alpha_6\pi - \alpha_2 V/P + [\alpha_2(1-u) - \alpha_3\lambda_3 + \alpha_5]\, B/P + (\alpha_4 + \alpha_3\lambda_2)\, H/P + g] \div [1 - \alpha_1 + \alpha_2 u + \alpha_3\lambda_1]$$

Assume that $V < 0$, $f > 0$, and the slope coefficients α_i, λ_i, and ρ are > 0, with $0 < \alpha_2 \leq \alpha_1 < 1$. We deduce that the denominator of ϕ in (7) is > 0, and that the partial derivatives ϕ_g, ϕ_H, and ϕ_π are > 0 while ϕ_u and ϕ_V are < 0. Assume $\lambda_2 H > \lambda_3 B$, i.e., that a price rise shifts the *LM* curve to the left via (4). This implies that $\phi_p < 0$. Also assume $\phi_H > r\phi_B$, i.e., that an open market purchase raises income in the short run via (7).

(b) Find the long-run equilibrium conditions for the system by setting all growth rates equal to their long-run equilibrium values. If a static equilibrium position is to be studied, then all growth rates are set to zero. If a dynamic steady-state growth path equilibrium is to be studied, then each growth rate is set equal to the value it is to have along that path.

For a static equilibrium in the above model, expected inflation π must be zero, and the static equilibrium conditions are like (1)–(6) except that (5) (after substituting for t from (2)) and (6) reduce, respectively, to

$$(5')\quad g - uy - V/P + (1-u)B/P = 0 \quad \text{(balanced budget)}$$

$$(6')\quad y = f \quad \text{(output = capacity)}$$

For a dynamic equilibrium with steady inflation at the rate $DP/P = \theta$ per year, but constant real variables such as y, g, H/P, etc.,

expected inflation π must be equal to θ, so that equation (6′) remains but (5′) becomes

$$(5'') \quad g - uy - V/P + (1-u)B/P = \theta\,(H/P + B/rP)$$

Note that in the inflationary equilibrium, condition (5″) provides not for a balanced budget, but for a constant real deficit, financed by the continual issue of bonds and high-powered money at the constant growth rate θ.

(c) Decide which policy variable is going to adjust endogenously. Notice that there is no need to make this decision until now, because it does not affect either the short-run solution of the static equations in step (a), or the set of long-run equilibrium conditions in step (b).

(d) Solve the long-run equilibrium conditions obtained in step (b) for the long-run equilibrium values of the endogenous variables, assuming given values of the exogenous variables. Long-run equilibrium multipliers can then be obtained by differentiating these solutions partially with respect to each exogenous variable. They will depend on the partial derivatives of ϕ found in step (a). Notice that the long-run equilibrium solutions, and the multipliers obtained by differentiating them, depend upon the decision as to which policy variable is to adjust endogenously.

The long-run equilibrium form of the foregoing model is block recursive. Equation (6′) alone shows that real income must equal the exogenous capacity level f in long-run equilibrium. The existence of any further recursiveness among the long-run equilibrium conditions depends on the choice of endogenous policy variable. For simplicity, consider static long-run equilibria and look at equations (7) and (5′).

Suppose first that high-powered money H is the endogenous policy variable. Then (5′) alone determines the long-run equilibrium price level:

$$(8) \qquad P = \frac{(1-u)B - V}{uf - g}$$

This equilibrium value is positive if and only if $uf - g$ is positive, that is, if and only if the constant marginal tax rate applied to capacity income yields more than enough revenue to cover government purchases g. When considering static equilibria we shall assume that this is the case. (For inflationary equilibrium paths, $uf - g$ need not be positive; of course the negative static equilibrium price level that results from (8) in such a case is of no practical interest. I shall return to this point later, in Section XI.) Once the equilibrium price level has been obtained from (8), the equilibrium high-powered money stock H can be obtained from (7). Suppose second that any other policy variable is endogenous instead of H. Then (5′) and (7) determine its long-run equilibrium value and that of P simultaneously, not recursively.

In an inflationary equilibrium, the equilibrium inflation rate θ must satisfy the *GBR* (5″).

Note that for any choice of endogenous policy variable, the long-run equilibrium solution and the associated multipliers are of no practical interest unless the long-run equilibrium is stable, because the path of the economy will not converge to an unstable equilibrium. If the static equilibrium price level is negative, it cannot be stable in a plausible model.

(e) Substitute the short-run equilibrium solutions found in step (a) into the dynamic equations of the model, thus obtaining a dynamic set of equations containing only those endogenous variables whose time derivatives appeared in the original model. This system will be different for different choices of endogenous policy variable.

(f) Linearize the dynamic system obtained in step (e) about its equilibrium. Then write the Routh-Hurwitz necessary and sufficient conditions for the stability of this system, and try to determine whether they are satisfied. In this model they involve the partial derivatives of ϕ found in step (a). Note that different stability conditions apply, depending on the choice of endogenous policy variable. Since there are as many of these conditions as there are dynamic equations in the model, it is sometimes not possible to determine whether all are satisfied, without empirical estimation. In such cases it is some-

times possible to discover theoretically that one of the conditions is definitely not satisfied, which proves instability.

(g) In unstable cases, examine the behavior of the system through time to see whether the unstable variables approach a constant-growth-rate path, such that certain ratios of the unstable variables (for example, real balances H/P) approach a constant value and thus are stable. This sometimes reveals that a system with an unstable price level has a stable rate of inflation and stable real variables.

I have noted that the choice of which policy variable is to adjust endogenously is important, because it affects the block recursiveness of the system, the long-run equilibrium multipliers, and the stability of the system. However, different choices of endogenous policy variable can lead the system to the same long-run equilibrium path, in the following sense: Any given long-run equilibrium path can be attained, from any initial position near that path, by any choice of endogenous policy variable for which the system is stable, provided that the paths chosen for the exogenous policy variables coincide with the given equilibrium path.

IV. Brief Comments on Selected *GBR* Models

Early work on the *GBR* (see my 1968 paper; Alan Blinder and Robert Solow; James Tobin and Willem Buiter, Sec. 3–5) used simple models in which the price level was assumed constant. This is unrealistic, for clearly one of the important effects of changes in the high-powered money stock, at least in some cases, is to change the price level. Later models treat prices as endogenous, and determine a long-run equilibrium price level. Some of these (see William Scarth; Tobin-Buiter, Sec. 6; my 1978 paper) treat productive capacity as fixed. This too is unrealistic, for capacity can be increased by growth in the capital stock. Several others allow for capacity growth as a result of capital accumulation (see Jürg Niehans, 1974, 1978, ch. 11; Tobin-Buiter, Sec. 6; Buiter, 1976, 1977; Karl Brunner and Allan Meltzer; Stephen Turnovsky, 1977, ch. 8; 1978). A few recent models analyze steady-inflation equilibria (see David Pyle and Turnovsky; David Currie, 1976; Turnovsky, 1977, ch. 8; my 1978 paper). A fev deal with steady-state real growth based on population growth (see Buiter, 1977; Turnovsky, 1978).

V. Five Types of Equilibrium

Five types of long-run equilibrium can usefully be distinguished:

(i) A static equilibrium position for all real and nominal variables. Here equilibrium requires a balanced budget. (See my 1968 paper; Blinder-Solow; Tobin-Buiter.)

(ii) A dynamic steady-state growth equilibrium path for all aggregate real variables, but a static equilibrium position for all aggregate nominal stock and flow variables. Here too equilibrium requires a balanced budget; it also requires a steadily declining price level so that real balances can grow in proportion to population while nominal balances are constant. (See Buiter, 1977.)

(iii) A dynamic steady-state growth equilibrium path for aggregate real variables, and a constant price level. This requires steady-state growth of nominal stocks and flows, and a constant real budget deficit.

(iv) A static equilibrium position for aggregate real variables, but a steady inflation. This requires steady-state growth in nominal stocks and flows, and a constant real budget deficit. (See Turnovsky, 1977, ch. 8; my 1978 paper.)

(v) A dynamic steady-state growth equilibrium path for aggregate real variables, and a steady inflation. This requires steady-state growth of nominal stocks and flows at a rate approximately the sum of the growth rates of real variables and prices. (See Turnovsky, 1978.)

Note that in equilibria of types (iii)–(v) where nominal stocks and flows grow steadily without limit, the system must fail to satisfy stability conditions pertaining to a static equilibrium of nominal aggregate levels. However, the dynamic equilibrium path may be stable. If so, it will satisfy stability conditions regarding real per capita stocks and flows.

TABLE 1—STABILITY CONDITIONS IN THE MODEL OF EQUATIONS (1)–(6) FOR VARIOUS CHOICES OF ENDOGENOUS POLICY VARIABLE

	Endogenous Policy Variable			
	V	g	H	B
Determinant appearing in long-run equilibrium multipliers and in stability conditions with sign if known	$\Delta_V = \frac{[(1-u)B-V]\phi_V}{P^2} - \frac{\phi_P}{P} > 0$	$\Delta_g = \frac{[(1-u)B-V]\phi_g}{P^2} + \phi_P$	$\Delta_H = \frac{[(1-u)B-V]\phi_H}{P^2} > 0$	$\Delta_B = \frac{[(1-u)B-V]\phi_B}{P^2} + \frac{(1-u)\phi_P}{P}$
Long-run equilibrium multiplier $\partial P/\partial H$	$(\phi_H/P)/\Delta_V$	$-\phi_H/\Delta_g$		$-[(1-u)\phi_H/P]/\Delta_B$
Necessary and sufficient conditions for local stability of static equilibrium[b]	$\frac{dDP}{dP} = \frac{-P^2\rho\Delta}{(1-\alpha_1+\alpha_3\lambda_1)f}\Delta_V < 0$	$\frac{dDP}{dP} = \frac{P\rho\Delta/f}{1-\alpha_1-(1-\alpha_2)u+\alpha_3\lambda_1}\Delta_g < 0$	$tr\ A_H = P(-u\phi_H + \rho\phi_P/f) < 0$; $det\ A_H = \frac{P^2\rho}{f}\Delta_H > 0$	$tr\ A_B = r(1-u-Pu\phi_B) + P\rho\phi_P/f < 0$; $det\ A_B = \frac{P^2\rho r}{f}\Delta_B > 0$
Determinantal stability condition restated in terms of Δ_i	$\Delta_V > 0$	$P\Delta_g < 0$[a]	$\Delta_H > 0$	$\Delta_B > 0$
Determinantal stability condition restated in terms of $\partial P/\partial H$	$P\frac{\partial P}{\partial H} > 0$	$P\frac{\partial P}{\partial H} > 0$		$P\frac{\partial P}{\partial H} > 0$
Is stability condition satisfied if the static equilibrium price level P is positive?	yes	yes iff $\partial P/\partial H > 0$	tr yes; det yes	tr uncertain; det no iff $\partial P/\partial H > 0$

[a]Assuming that $[1-\alpha_1-(1-\alpha_2)u+\alpha_3\lambda_1] > 0$
[b]tr denotes the trace and det the determinant value

VI. Stability and Multipliers in the Model of Section II

I now review the stability conditions and long-run price multipliers derived in my 1978 paper for the foregoing model, equations (1)–(6). This will illustrate some of the ideas mentioned above, and also prepare the way for a partial solution of the instability problem that appears in some cases when bonds are endogenous.

Table 1 presents the results. Each column of the table corresponds to a different choice of endogenous policy variable. The first row presents important determinants, which (according to Samuelson's correspondence principle) appear both in the stability conditions and in the long-run equilibrium multipliers. The second row presents the equilibrium multipliers for the effect of the high-powered money stock on the price level, $\partial P/\partial H$. The denominator of each is the determinant in the first row. The numerator of each depends on the positive partial derivative ϕ_H from equation (7). (The second row has no entry when H is endogenous, of course.) The third row

presents the necessary and sufficient conditions for stability of a static equilibrium. If V or g is endogenous, there is just one dynamic equation, based on (6), and hence one necessary and sufficient stability condition; see the columns labeled V and g. When H or B is endogenous, there are two dynamic equations, based on (5) and (6), and hence two necessary and sufficient conditions: the characteristic matrix of the dynamic system must have a negative trace and a positive determinant; see columns H and B. The fourth and fifth rows restate the determinantal stability conditions from the third row in terms of the sign of the determinant shown in the first row, and in terms of the sign of the equilibrium multiplier $\partial P/\partial H$, respectively. The last row indicates whether each stability condition is satisfied under the assumptions made in Section III, step (a) above, and the assumption that the static equlibrium price level is positive, i.e., that $uf > g$, recall equation (8). Then we see in the last row that the system is unambiguously stable when either V or H is chosen to be the endogenous policy variable, but stability is related to the sign of the equilibrium multiplier $\partial P/\partial H$ when either g or B is made endogenous. In particular, when g is endogenous a positive sign for $\partial P/\partial H$ is necessary and sufficient for stability, but when B is endogenous a positive sign for $\partial P/\partial H$ implies instability. Ettore Infante and Jerome Stein, p. 492, found the same result when B is endogenous.

In the cases where g or B is endogenous, can we find plausible conjectures about stability and about the signs of the long-run equilibrium multipliers $\partial P/\partial H$? Or must these questions be left for empirical estimation of parameters? Historical evidence suggests that the effect of the money stock on prices is positive, not negative, and that the economy does not run away to plus or minus infinity when the money stock and tax transfer rates are held fixed. This suggests that $\partial P/\partial H$ is positive and the economy is stable. We have seen that in this model when g is endogenous these two statements are compatible (indeed, they are equivalent) so we might conjecture they are correct, but when B is endogenous they are incompatible.

Three conjectures are then open for the case of endogenous bonds: 1) $\partial P/\partial H$ is positive and the system is unstable; 2) $\partial P/\partial H$ is negative and the system is stable; or 3) any model that requires a choice between 1) and 2) is incorrect. History has not performed the crucial experiment that would decide among these three, namely, starting from an equilibrium, increase the money stock by an open market operation and forever afterward hold constant the money stock and all other policy variables except B. If the model be accepted, I am inclined to believe in instability when B is endogenous, partly because of doubt that $\partial P/\partial H$ is negative in a stable system, and partly because of a different stability argument proposed by Niehans (1977), to which I now turn.

Niehans' argument is this. Assume that the government budget restraint is always kept in equilibrium, but that output responds to aggregate demand with a lag. Postulate that this dynamic system is stable (this postulate is widely accepted). Note the resulting restrictions on the parameters of the *IS* and *LM* curves. Then note the implications of those restrictions for the stability of the original system. Niehans applied this argument to the first (i.e., the fixed-price) model in Blinder and Solow; he found that it restricts their *IS* and *LM* curves so that their system is unstable when bonds are endogenous. I have applied it to the model consisting of equations (1)–(6), assuming that the price level adjusts to equilibrium instantly. It implies restrictions on the *IS* and *LM* parameters, which in turn require that $\Delta_B < 0$ in Table 1, that is, that when bonds are endogenous in (1)–(6) the system is unstable and $\partial P/\partial H$ is positive. (The same argument establishes stability when g is endogenous, confirming the conjecture above.)

Note that when H is endogenous, the discriminant $tr^2 A_H - 4 \; det \; A_H$ can be either negative or positive, so cycles may or may not occur. When B is endogenous, if Δ_B is negative, the discriminant is positive, so there are no cycles.

VII. Instability when Bonds are Endogenous

Several other writers in the *GBR* literature have found a similar theoretical result,

namely that stability is impossible (or unlikely) when bonds are endogenous, but certain (or likely) when money is endogenous. This result occurs in studies that use government purchases g as the government expenditure variable, including Blinder-Solow, Infante-Stein, and my 1978 paper as noted above, and Scarth and Pyle-Turnovsky. It also occurs in Tobin-Buiter (Sec. 6) and Buiter (1976, 1977) which use, as the government expenditure variable, government purchases plus government interest net of tax, denoted by g':

$$g' = g + (1 - u)B/P \tag{9}$$

The policy that is unstable (or likely to be so) in these studies resembles a policy that has been advocated by some of the monetarists, that is, fix government purchases and tax-transfer schedules, make the money stock grow at a constant moderate rate, and allow any variation in deficits or surpluses to be covered by issuing or retiring government debt. For example, Milton Friedman says,

> The right policy—not alone for this episode but as a general rule—is to *let the quantity of money increase at a rate that can be maintained indefinitely without inflation* (about 5 per cent per year) and to keep taxes and spending at levels that will *balance the budget at high employment*. [p. 92]

It is not known whether the *U.S.* economy has stability properties like the models in these studies, but if it does, a policy of constant moderate growth of the money stock with fixed g or g', fixed tax-transfer rates, and endogenous bonds would appear to have stability problems.

The difficulty appears to be associated with the fact that when bonds are issued to finance a deficit, the deficit is not thereby reduced. First, if g is the expenditure variable held constant during the disequilibrium, the *GBR* equation (5) with taxes given by (2) shows that when the debt rises to cover a deficit, the deficit *increases* because of the increase in the term $(1 - u)B/P$ which represents interest payments net of tax. Thus the system is moved further from equilibrium, rather than toward it. Second, if g' is the expenditure variable held constant, the term in B/P disappears from the *GBR*, which becomes

$$g' - uy - V/P = DH/P + DB/rP \tag{10}$$

Hence the deficit *does not change* as more bonds are issued, and so the system is not moved toward equilibrium.

VIII. Some Hope for Stability when Bonds are Endogenous

This suggests that a solution to the instability problem when bonds are endogenous might be to use a different government expenditure variable, namely, government purchases plus debt interest *gross* of tax, denoted by g'':

$$g'' = g + B/P \tag{11}$$

(This suggestion was made independently by Frank De Leeuw in correspondence.) The *GBR* (5) then becomes the following:

$$g'' - uy - V/P - uB/P = DH/P + DB/rP \tag{12}$$

In this case, when bonds are issued to finance a deficit with g'' constant, the term $(-uB/P)$ decreases algebraically, which *decreases* the deficit and moves the system toward equilibrium, *ceteris paribus*. What causes the difference, of course, is that any increase in debt interest must be counteracted by an equal decrease in government purchases, g.

Let us proceed to analyze the stability of the model when g'' is used as the expenditure variable rather than g. It is first necessary to replace g by $g'' - B/P$ not only in the original *GBR* (5) to obtain (12), but also in (1), (5'), and (5''). We shall make the same assumptions as in Section III, step (a) above, after replacing g by $g'' - B/P$ in the numerator of ϕ in (7). We shall again assume the equilibrium price level is positive; this is equivalent to $uf > g$ or g' or g'' in the three models, respectively. And we shall assume that $-uB - V > 0$, i.e., a price rise moves the budget toward surplus; see equation (12).

Table 2 summarizes the results for the model using g'' in the third row. For comparison, it also shows in the first and second rows

TABLE 2—STABILITY OF THE MODELS FOR THREE DIFFERENT GOVERNMENT EXPENDITURE VARIABLES AND VARIOUS CHOICES OF ENDOGENOUS POLICY VARIABLE

Government Expenditure Variable	Endogenous Policy Variable			
	V	g or g' or g''	H	B
g = Purchases of Goods and Services	Stable	Stable iff $\partial P/\partial H > 0$	Stable	Unstable if $\partial P/\partial H > 0$, because the condition that $det\ A_B > 0$ is equivalent to $\partial P/\partial H < 0$
$g' = g$ + Interest Net of Tax $= g + (1 - u)B/P$	Stable iff $\partial P/\partial H > 0$	Stable iff $\partial P/\partial H > 0$	Stable	Unstable if $\partial P/\partial H > 0$, because the condition that $det\ A_B$ be > 0 implies that $\partial P/\partial H = 0$[a]
$g'' = g$ + Interest Gross of Tax $= g + B/P$	Stable iff $\partial P/\partial H > 0$	Stable iff $\partial P/\partial H > 0$	Stable	Possibly stable: the condition that $det\ A_B$ be > 0 is satisfied iff $\partial P/\partial H > 0$, but the condition that $tr\ A_B$ be < 0 is uncertain

Notes: See Table 1. The stability conditions above are obtained under the assumption that the equilibrium price level is positive. When g' or g'' is used instead of g, the expression $1 - u$ that appears in Table 1, in the four determinants in the first row above, and in the expressions for $\partial P/\partial H$ and $tr\ A_B$ in column B, is replaced respectively by zero or by $-u$.

[a]Here $\partial P/\partial H \neq 0$ iff $det\ A_B = 0$. In cases like this where the linearized system is unstable because a zero value appears where stability would require a positive one, it sometimes happens that the original non-linear system is stable

the results for the original model using g (taken from Table 1, last row), and for the model that uses g'. The third column shows that when H is endogenous, all three models are stable. The second column shows that when government expenditure is endogenous, stability in all three models is equivalent to a positive sign for $\partial P/\partial H$. The first column shows that when V is endogenous, the shift from g to either g' or g'' makes stability equivalent to a positive sign for $\partial P/\partial H$. And the fourth column shows that when B is endogenous, the shift from either g or g' to g'' reverses the previous relation between the sign of $\partial P/\partial H$ and the sign of the stability determinant: a positive sign for $\partial P/\partial H$ is now equivalent to a *positive* sign of that determinant, so that the system may now be stable, depending upon the trace condition (whose satisfaction is uncertain under our assumptions).

Suppose the first model in Blinder-Solow is modified to use g'', government purchases plus gross interest, as the expenditure variable. Then a similar argument shows that the model remains stable when money is endogenous, and that when bonds are endogenous a positive sign for the equilibrium multiplier $\partial Y/\partial H$ is equivalent to stability, where Y is their income variable. Further, the Niehans (1977) argument applied to the Blinder-Solow model thus modified also indicates that it is definitely stable when either bonds or money is endogenous.

Tobin-Buiter (Sec. 6) found that their model with static expectations is definitely unstable when bonds are endogenous because the determinant of the characteristic equation is unambiguously negative. When that model is modified to use government purchases plus gross interest (rather than net interest, which they use), the sign of the determinant of the characteristic equation becomes uncertain so that stability is now possible.

We have found that in three models (see my 1978 paper; Blinder-Solow, first model; Tobin-Buiter, Sec. 6, static-expectations model), the use of g'', government purchases plus gross interest, as the expenditure variable weakens the conclusion that the stability determinants are negative and that the models are therefore unstable when bonds are

endogenous, though it does not definitely establish stability except in the first Blinder-Solow model. Thus the above mentioned monetarist policy of fixing a moderate constant growth rate of the money stock with bonds as the endogenous policy variable is not yet completely rescued from the demon of instability, but it is given some hope.

IX. When is the Multiplier Equal to the Inverse of the Tax Rate?

A rather common result in the *GBR* literature is that the equilibrium multiplier effect of government expenditure on income is simply the inverse of marginal tax rate, $1/u$. This is a striking result, for the behavior parameters of the economy have no effect on it. However, it is unrealistically special. My 1968 paper and Blinder-Solow obtain it when the endogenous policy variable is high-powered money (but not when it is bonds) in a simple model where the price level is rigid, there is excess capacity, and the government expenditure variable is government purchases. Of course, such a result makes no practical sense except temporarily in a deep depression, for in reality prices are not rigid and capacity limits output.

Tobin-Buiter (Sec. 6) and Buiter (1976, 1977) obtain the same multiplier, $1/u$, when the endogenous policy variable is bonds in a full-employment model where prices and capacity are endogenous, bonds are short term, the tax-transfer system is proportional ($V = 0$, in my notation), and the government expenditure variable is government purchases plus interest net of tax (g' in my notation). Buiter (1976) shows that the same multiplier arises in his model when the endogenous variable is high-powered money. In fact the same is true of the Tobin-Buiter model, as is easily shown.

In general, when prices are endogenous and capacity limits output, the equilibrium multiplier effect of government expenditure on income is not equal to the inverse of the marginal tax rate. Two special features of the Tobin-Buiter and Buiter models combine to produce this special result. One is the assumption of a proportional tax system ($V = 0$). The other is the use of g', government purchases plus net interest, as the expenditure variable. With these two modifications, the static-equilibrium balanced-budget version of the *GBR* in (5′) becomes

$$g' - uy = 0 \tag{13}$$

From this it is clear that the multiplier $\partial y/\partial g'$ equals $1/u$ regardless of whether the endogenous policy variable is bonds or high-powered money. If a nonzero intercept V is introduced into their tax-transfer function, the multipliers $\partial Y/\partial g'$ for endogenous bonds and for endogenous high-powered money are no longer equal to each other or to $1/u$. The same occurs if the government expenditure variable g' is replaced by g or by g''.

X. When are Money-Financed Expenditures More Powerful than Bond-Financed Expenditures?

Blinder-Solow and Pyle-Turnovsky have inquired whether bond-financed expenditure is more expansionary than money-financed expenditure, and have concluded, respectively, that the answer is yes or maybe. They did so by comparing the long-run equilibrium multipliers for the two cases when the endogenous policy variable is bonds or high-powered money. Blinder-Solow compared multiplier effects on income (recall they assumed prices fixed), while Pyle-Turnovsky compared multiplier effects on the inflation rate. This method is appropriate only if the system is stable under both endogenous bonds and endogenous money so that when disturbed, it does converge to the state that is described by the multipliers. However, we have seen reason to believe that the Blinder-Solow first model is unstable when bonds are endogenous. If so, the equilibrium multiplier for bond finance (though inappropriate because of instability) is less than $1/u$, the equilibrium multiplier for money finance. The Pyle-Turnovsky equilibrium multipliers may also be inappropriate because they find that their system may be unstable under either bond finance or money finance, more likely so under bond finance.

When the Blinder-Solow first model is modified to use g'', government purchases plus gross interest, as the expenditure variable, and when the multiplier $\partial Y/\partial H$ is

positive, the model is stable and the multiplier for bond finance (now appropriate because of stability) becomes less than $1/u$, the multiplier for the money finance. Thus in this case money finance is unambiguously more powerful than bond finance. When the model in my 1978 paper is modified in the same way, the price multiplier $\partial P/\partial g''$ for bond finance (now appropriate if bond finance is stable, which it may be) also becomes less than the price multiplier $\partial P/\partial g''$ for money finance. Thus also in this case money finance is more powerful than bond finance if the latter is stable.

XI. Instability in Nominal Terms is Compatible with Stability in Real Terms

Consider the original model of equations (1)–(6) again, but now assume (as promised earlier) that government purchases exceed the revenue brought in by the marginal tax transfer rate applied to capacity income, so that $g > uf$. Then, as noted earlier, the equilibrium price level in (8) becomes negative, and is irrelevant for the economy's actual behavior.

Consider the case where high-powered money H is endogenous, with bonds B and nominal taxes less transfers V fixed. Then the determinant condition for stability is satisfied, but the trace condition is violated at the negative equilibrium price, which is therefore unstable (recall Table 1, column H). There is a continual deficit which is financed by issuing money. The price level rises without limit as does the money stock. The real values of bond interest B/P and autonomous taxes less transfers V/P go to zero as P goes to infinity. Thus the system approaches one in which taxes less transfers are strictly proportional to income and there is no government debt.

There may be a stable inflationary equilibrium path along which real balances H/P and the inflation rate θ are constant and the constant real deficit $g - uy$ is financed by the so-called inflation tax yielding a real "revenue" of $\theta H/P$. In that case the inflationary-equilibrium GBR in (5″) becomes

$$g - uy = \theta H/P \tag{14}$$

This situation can occur with the expected inflation rate π equal to the actual rate θ, provided that the deficit is small enough so that there exists a pair of noncomplex values of the inflation rate and real balances demanded such that their product $\theta H/P$ can equal the deficit. If the deficit is too large, it cannot be financed by issuing money at a steady rate because the attempt to do so will create such a rapid inflation that real balances, which are the taxable base of the inflation tax, are driven down too low. Mathematically, we are seeking a simultaneous solution to equations (14) and (7) for the two variables θ and H/P, after modifying (7) by dropping out the terms in V/P and B/P (which go to zero), replacing income y by capacity f and replacing the expected inflation rate π by the actual rate θ. There are two solutions because (14) is a hyperbola while (7) is linear. Suppose (7) traverses the first quadrant of the $(\theta, H/P)$ space. It slopes negatively. The lines intersect twice in the first quadrant, or have a single tangency point there, or else don't intersect at all. If they don't intersect, the solutions are complex numbers and there is no equilibrium for θ and H/P in the domain of real numbers. If they do intersect, both solutions are real and positive. Given the expected inflation rate, only one solution can be attained. It is the one whose inflation rate is expected. The stability analysis is simplified because the two dynamic variables DH and DP can be combined into one, $D(H/P)$, so that there is only a single dynamic equation in the system. Under our assumptions it is easy to show that this equilibrium is stable.

A similar situation arises if the deficit is financed not by money alone, but by money and bonds in a fixed proportion, except that then the real debt approaches a positive equilibrium value rather than zero. In each case there is no stable static equilibrium for the price level and nominal income and asset holdings, but there may be one for the inflation rate and real income and asset holdings.

A parallel situation also arises in the presence of steady-state real aggregate growth, where there is no stable equilibrium for aggregate real income, but there may be one for per capita real income.

XII. *GBR* Models Provide an Impartial View of the Monetarist-Fiscalist Controversy

The theory of *GBR* models does not require the extreme monetarist conclusion that the effect of a macro-economic policy change depends solely upon what happens to the stock of money. Nor does it require the extreme fiscalist conclusion that the effect of a macro-economic policy change is independent of what happens to the stock of money. It suggests that changes in the money stock, the debt, government purchases, transfers, and taxes all have effects, depending on which variables are fixed and which are allowed to vary endogenously.

XIII. Limitations

This paper is subject to many limitations. There is no foreign sector. There is no banking sector and no conventional quantity of money. Dynamic systems of at most two differential equations are considered. Endogenous short-run expectations about prices and inflation are not considered. Steady-state real growth paths are barely mentioned. There are no lags in the adjustment of output to aggregate demand. The Phillips curve is assumed vertical in the long run. Empirical problems and results are not considered. There is no discussion of discretionary stabilization policy or of policy optimization.

REFERENCES

A. S. Blinder and R. M. Solow, "Does Fiscal Policy Matter?," *J. Publ. Econ.*, Nov. 1973, *2*, 319–37; "Does Fiscal Policy Matter? A Correction," *J. Publ. Econ.*, Jan./Feb. 1976, *5*, 183–84.

K. Brunner and A. H. Meltzer, "An Aggregate Theory for a Closed Economy," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976, 69–103.

W. H. Buiter, "Capacity Constraints, Government Financing and the Short Run and Long Run Effects of Fiscal Policy," mimeo., Princeton Univ. 1976.

———, "An Integration of Short Run Neo-Keynesian Analysis and Growth Theory," *De Economist*, No. 3, 1977, *125*, 340–59.

C. F. Christ, "A Simple Macroeconomic Model with a Government Budget Restraint," *J. Polit. Econ.*, Jan./Feb. 1968, *76*, 53–67.

———, "Some Dynamic Theory of Macroeconomic Policy Effects on Income and Prices under the Government Budget Restraint," *J. Monet. Econ.*, Jan. 1978, *4*, 45–70.

D. A. Currie, "The Government Budget Constraint and Inflation: Some Dynamic Implications," disc. paper, Queen Mary College, London 1976.

———, "Macroeconomic Policy and Government Financing," Association of University Teachers of Economics *Proceedings*, 1977.

Milton Friedman, *An Economist's Protest*, Glen Ridge 1972, 41; reprinted from *Newsweek*, October 17, 1966, p. 92.

E. F. Infante and J. L. Stein, "Does Fiscal Policy Matter?," *J. Monet. Econ.*, Nov. 1976, *2*, 473–500.

Jürg Niehans, "Monetary Policy with Full Stock Adjustment," *Z. Wirtsch. Sozialwiss.*, No.1, 1974, 17–43.

———, "A Comment on Stabilization Paradoxes," in Horst Albach, Ernst Helmstadter and Rudolf Henn, eds., *Quantitative Wirtschaftsforschung, Wilhelm Krelle zum 60. Geburtstag*, Tubingen 1977, 513–21.

———, *The Theory of Money*, Baltimore 1978.

D. H. Pyle and S. J. Turnovsky, "The Dynamics of Government Policy in an Inflationary Economy," *J. Money, Credit, Banking*, Nov. 1976, *8*, 411–37.

W. M. Scarth, "The Effects on Economic Stability of Indexing the Tax System and Government Bond Yields for Inflation," *Can. J. Econ.*, Aug. 1975, *8*, 383–98.

J. Tobin and W. H. Buiter, "Long-Run Effects of Fiscal and Monetary Policy on Aggregate Demand," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976, 273–309.

Stephen J. Turnovsky, *Macroeconomic Analysis and Stabilization Policy*, Cambridge 1977.

———, "Macroeconomic Dynamics and Growth in a Monetary Economy," *J. Money, Credit, Banking*, Feb. 1978, *10*, 1–26.

An Analysis of a Macro-Econometric Model with Rational Expectations in the Bond and Stock Markets

By RAY C. FAIR*

The difficulty of accounting for expectational effects in macro-economic models is well known. The standard procedure in dealing with this problem in the construction of large-scale macro-econometric models is to use current and lagged values as "proxies" for expected future values. An alternative procedure is to assume that expectations are rational. Although the assumption of rational expectations has received increased attention lately in work with theoretical and small-scale empirical models,[1] it has not yet been applied to large-scale macro-econometric models. In this paper the assumption that expectations are rational in the bond and stock markets will be applied to a large-scale macro-econometric model. The quantitative effects of monetary and fiscal policies in this model will be compared to those in a similar model without rational expectations. The quantitative sensitivity of monetary and fiscal policy effects to alternative expectational assumptions is clearly an important question in macroeconomics, and the primary purpose of this paper is to provide an estimate of this sensitivity for the assumption of rational expectations in the bond and stock markets.

The econometric model that is used in this study is the one in my 1976 book. Three "versions" of this model are analyzed: the original version and two modified versions. The original version, which will be called Model 1, does not have rational expectations in the bond and stock markets. There are two term-structure equations and one stock-price equation in the model, and in these three equations current and lagged values are used as proxies for expected future values. In the first modified version, which will be called Model 2, the two term-structure equations are replaced with a specification that is consistent with the existence of rational expectations in the bond market.[2] The second modified version, which will be called Model 3, is the same as Model 2 except that the stock-price equation is replaced with a specification that is consistent with the existence of rational expectations in the stock market. In Model 3, therefore, there are rational expectations in both the bond and stock markets, and it is to my knowledge the first example of a large-scale econometric model for which this is true.

It is important to note at the outset that this paper does not contain a test of the assumption of rational expectations in the bond and stock markets. Within the context of the present model it is more difficult to test this assumption than it is to examine its policy implications, and such a test is beyond the scope of the present paper. The way in which this assumption could be tested using the present model is discussed in footnote 13 below. It is also important to note at the outset that Model 3 is not a model in which all expectations are rational. Model 1 contains

*Cowles Foundation, department of economics, Yale University. The research described in this paper was financed by grant SOC77-03274 from the National Science Foundation. This paper is a revised version of a paper presented at the Econometric Society Meetings, Ottawa, June 22, 1977.

[1]For one class of models with rational expectations, see Robert E. Lucas, Jr., Thomas J. Sargent (1973, 1976), Sargent and Neil Wallace, and Robert J. Barro. See also my 1978b paper for a criticism of this class of models. For an example of the use of the assumption of rational expectations in a small-scale empirical model (the St. Louis model), see Paul A. Anderson.

[2]The idea of replacing term-structure equations in macro-econometric models with a specification that is consistent with the existence of rational expectations in the bond market is contained in a paper by William Poole (pp. 477–78). Poole (p. 478) questioned the computational feasibility of this procedure for large-scale models, but, as discussed below, this procedure is in fact computationally feasible.

markets other than the bond and stock markets, such as the labor and goods markets, in which expectations are not rational, and Model 3 differs from Model 1 only with respect to the bond and stock markets. It is again beyond the scope of the present paper to consider an econometric model in which all expectations are rational.

The outline of this paper is as follows. A few of the features of the original version of the model are reviewed in Section I, and then the modifications of it are considered in Section II. The basic experiments that were performed using the three versions are described in Section III, and the results of these experiments are presented and discussed in Section IV. Some further experiments and results are described in Section V. Section VI contains a brief summary of the main conclusions of this study.

I. A Brief Review of Model 1

Model 1 consists of eighty-four equations, twenty-six of which are stochastic. There are five sectors (household, firm, financial, foreign, and government) and five categories of financial securities (demand deposits and currency, bank reserves, member bank borrowing from the Federal Reserve, gold and foreign exchange, and an "all other" category). Since the model is described in detail in my 1976 book, no extensive discussion of it will be presented here. It will be useful for purposes of the following analysis, however, to review briefly the interest rate and wealth effects in the model and the structure of the financial sector.

There are three endogenous interest rates in the model: the three-month Treasury Bill rate (r), an Aaa corporate bond rate (RA), and a mortgage rate (RM);[3] the last an explanatory variable in three of the four consumption equations and in one of the three labor supply equations. RA is an explanatory variable in the two interest payment equations, in the stock-price equation, and in the main price equation of the model; and r is an explanatory variable in two of the consumption equations, in the two demand-for-money equations, in an equation explaining member bank borrowing from the Federal Reserve, and in one of the interest payment equations. In addition, there is a loan-constraint variable in the model that is a function of r, and this variable is an explanatory variable in one of the consumption equations and in an equation explaining the dividend payments of the firm sector; r is also an explanatory variable in the term-structure equations for RA and RM.

The variable explained by the stock-price equation is CG, the value of capital gains (+) or losses (−) on corporate stocks held by the household sector. The variable CG is part of the definition of ΔA in the model, where A is the value of securities (other than demand deposits and currency) held by the household sector. I use A lagged one-quarter as an explanatory variable in three of the consumption equations and in one of the labor supply equations.

A key feature of the model regarding the financial sector is that it is closed with respect to the flows of funds in the system. This means that any financial saving or dissaving of a sector in a period results in the change in at least one of its assets or liabilities, and that a financial asset of one sector is a corresponding liability of some other sector. It also means that the government budget constraint is explicitly accounted for and that the amount of government securities outstanding can be taken to be a direct policy variable of the Federal Reserve (henceforth called the Fed). The government budget constraint is

$$(1)\quad 0 = SAVG_t + \Delta VBG_t + \Delta(BR_t - BORR_t) + EXOG_t$$

where $SAVG$ is the financial saving of the government sector (the negative of the government deficit), VBG is the amount of government securities outstanding, BR is the amount of bank reserves, $BORR$ is the amount of member bank borrowing from the Fed, and $EXOG$ denotes the remaining variables in the equation, all of which are exogenous. The terms $SAVG$, BR, and $BORR$ are endogenous variables in the model and are explained by other equations. Equation (1) states that any nonzero level of saving of the government must result in the change in

[3]In terms of the notation in my 1976 book, r = $RBILL$, RA = $RAAA$, and RM = $RMORT$.

either VBG or nonborrowed reserves. Government securities are included in the "all other" category of securities in the model, and there is an equation in the model that equates the aggregate supply of this category to the aggregate demand. This equation can be written:

$$(2) \qquad 0 = \Delta VBG_t - \Delta VBP_t$$

where VBP is the amount of government securities held by the nongovernment sectors.[4] It is also an endogenous variable and determined elsewhere in the model.

The government budget constraint (1) is redundant, and so it can be dropped from the model. This still leaves equation (2), however, as an "extra" equation. Since VBP is determined elsewhere in the model, if VBG is taken to be exogenous, then some variable not explained by any other equation must be chosen to be endogenous in order to close the model. The variable chosen in this case is the bill rate (r). There is thus no equation in the model in which the bill rate appears naturally on the left-hand side. The bill rate is instead implicitly determined: its solution value each period is (speaking loosely) the value that makes equation (2) hold. As noted above, r is an explanatory variable in a number of the key equations in the model.

If, as just discussed, VBG is taken to be exogenous, then the behavior of the Fed is exogenous. In other words, the behavior of the Fed is not influenced by the state of the economy. In a recent study (1978a) I have estimated an equation explaining Fed behavior, and since this equation is used for some of the experiments below, it will be useful to provide a brief review of it. In this study the Fed was assumed to choose each period an optimal value of the bill rate and then to achieve this value through changes in its three policy variables: the reserve requirement ratio (g_1), the discount rate (RD), and VBG. Based on this assumption, an equation explaining the bill rate was estimated, where the explanatory variables were taken to be variables that seemed likely to affect the Fed's optimal value of the bill rate. This estimated equation was then interpreted as an explanation of Fed behavior. The equation is[5]

$$\begin{aligned}(3) \quad r_t = & \underset{(2.93)}{-11.1} + \underset{(16.30)}{0.841 r_{t-1}} \\ & + \underset{(1.69)}{0.0497\% PD_{t-1}} + \underset{(2.97)}{0.0352 J^*_t} \\ & + \underset{(1.62)}{0.0427\% GNPR_t} + \underset{(1.36)}{0.0188\% GNPR_{t-1}} \\ & + \underset{(2.10)}{0.0251\% M_{1t-1}}\end{aligned}$$

$$\hat{\rho} = \underset{(2.28)}{0.229}, \; S.E. = 0.474,$$

$$R^2 = 0.939, \; D.W. = 1.82$$

Sample Period = 1954I–1976II

where $\%PD$ is the percentage change at an annual rate in the price deflator for domestic sales, J^* is a measure of labor market tightness, $\%GNPR$ is the percentage change at annual rate in real GNP, $\%M_1$ is the percentage change at an annual rate in the money supply, and $\hat{\rho}$ is the estimate of the first-order serial correlation coefficient. The t-statistics in absolute value are in parentheses. Equation (3) states that the current bill rate is a positive function of the lagged rate of inflation, of the current degree of labor market tightness, of the current and lagged rates of growth of real GNP, and of the lagged rate of growth of the money supply. The behavior reflected in this equation is thus behavior in which the Fed "leans against the wind." The wind in this case is composed of the inflation rate, the degree of labor market tightness, the growth rate of real GNP, and the growth rate of the money supply. As these variables rise, so also does the bill rate.

If equation (3) is added to the model, then the behavior of the Fed is endogenous. In this case one of the three policy variables of the Fed (g_1, RD, or VBG) must be taken to be endogenous in order to close the model. For the experiments below VBG was always chosen to be the endogneous variable in this

[4]Equation (2) is equation 70 in my 1976 book. For purposes of the discussion in this paper, the original notation has been simplified: ΔVBP_t represents all the terms in equation 70 except ΔVBG_t.

[5]Equation (3) was estimated under the assumption of first-order serial correlation of the error term by the two-stage least squares technique described in my 1970 paper. The two endogenous explanatory variables in the equation are J^*_t and $\%GNPR_t$.

case. Thus the solution value for *VBG* each period is (again speaking loosely) the value that makes equation (2) hold.

To summarize this review of the financial sector of Model 1, r affects directly *RA*, *RM*, the loan-constraint variable, two consumption variables, two demand-for-money variables, member bank borrowing from the Fed, and one interest payment variable. It affects indirectly through the loan-constraint variable one consumption variable and one dividend variable. It affects indirectly through *RA* and *RM* three consumption variables, one labor supply variable, two interest payment variables, the main price variable in the model (*PF*), and *CG*. In addition, *CG* affects *A*, which affects with a lag of one-quarter three consumption variables and one labor supply variable. These latter effects are wealth effects on the household sector. The variable *PF* affects the level of sales, which affects the level of production, which affects the levels of investment and employment. The bill rate r thus has an indirect effect on investment and employment through its indirect effect on *PF* and in turn is affected by all the other variables in the model when it is implicitly determined (*VBG* exogenous). When the behavior of the Fed is endogenous (*VBG* endogenous), then r is determined according to the Fed behavioral equation.

II. The Modifications of Model 1

A. *The Term Structure Equations*

In order to consider the policy implications of the assumption of rational expectations in the bond market, some assumption about the determination of the term structure of interest rates must first be made. The following analysis is based on the assumption that the term structure is determined according to the expectations theory. Although this theory as applied below abstracts from considerations of such things as transactions costs and preferred habitats, the following analysis could be easily modified to incorporate a slightly different theory. What is needed for the work below is some link between long rates and expected future short rates, not that this link necessarily be the one postulated by the expectations theory.

According to the expectations theory of the term structure of interest rates (which should not be confused with the assumption of rational expectations), the return from holding an n-period security is equal to the expected return from holding a series of one-period securities over the n periods. Let ${}_tr^e_{t+i}$ denote the expected one-period rate of return for period $t + i$, the expectation being conditional on information available as of the beginning of period t, and let R_t denote the yield to maturity in period t on an n-period security. Then according to the expectations theory:

$$(4) \quad (1 + R_t)^n = (1 + {}_tr^e_t)(1 + {}_tr^e_{t+1}) \dots (1 + {}_tr^e_{t+n-1})$$

Since the ${}_tr^e_{t+i}$ values in equation (4) are unobserved, some assumption about how expectations are formed must be made in order to implement this theory. Model 1 rests on the assumption that each ${}_tr^e_{t+i}$ is a function of r_t, of lagged values of r_t, and of lagged values of the inflation rate.[6] Given this assumption, R_t is then according to equation (4) a function of these same variables. The estimated equations for RA_t and RM_t in Model 1 that are meant to approximate this function are[7]

[6]The possibility that both lagged values of the nominal rate and lagged values of the inflation rate affect expectations of future nominal rates is discussed by Franco Modigliani and Robert J. Shiller, pp. 19–23.

[7]For purposes of the work in this paper the model in my 1976 book was reestimated through 1976II using the revised national income accounts data. The estimated coefficients in equations (5) and (6) thus differ somewhat from those presented in my 1976 book, Table 2-3 Likewise, the estimates presented in equation (10) below for *CG* differ somewhat from the original estimates. The estimated version of the model used in this study is the same as the one used for the results in my 1978a paper. The $\hat{\rho}$ in equation (6) is the estimate of the first-order serial correlation coefficient. The t-statistics in absolute value are in parentheses. Equation (6) was estimated under the assumption of first-order serial correlation of the error term by the technique discussed in my 1970 paper. Equations (5) and (10) were estimated by the standard two-stage least squares technique. The endogenous explanatory variables are r_t in equations (5) and (6) and RA_t and Π_t in equation (10).

$$(5)\ \log RA_t = \underset{(4.10)}{0.0695} + \underset{(46.86)}{0.915}\log RA_{t-1} + \underset{(3.07)}{0.1767}\log r_t + \underset{(3.07)}{0.1867}\log r_{t-1} + \underset{(2.34)}{0.0636}\log r_{t-2} + \underset{(2.23)}{1.27}\left(\frac{1}{2}\Delta\log PX_{t-1} + \frac{1}{3}\Delta\log PX_{t-2} + \frac{1}{6}\Delta\log PX_{t-3}\right)$$

$R^2 = 0.996$, $S.E. = 0.0223$, $D.W. = 1.80$

Sample Period = 1954I–1976II

$$(6)\ \log RM_t = \underset{(3.87)}{0.1965} + \underset{(24.33)}{0.852}\log RM_{t-1} + \underset{(0.70)}{0.0297}\log r_t + \underset{(1.56)}{0.0854}\log r_{t-1} + \underset{(3.00)}{0.1138}\log r_{t-2} + \underset{(2.38)}{0.0551}\log r_{t-3} + \underset{(1.87)}{1.59}\left(\frac{1}{2}\Delta\log PX_{t-1} + \frac{1}{3}\Delta\log PX_{t-2} + \frac{1}{6}\Delta\log PX_{t-3}\right)$$

$\hat{\rho} = \underset{(2.41)}{0.247}$, $R^2 = 0.988$,

$S.E. = 0.0254$, $D.W. = 1.93$

Sample Period = 1954I–1976II

The last term in equations (5) and (6) is a weighted average of the rates of inflation in the past three quarters, with weights of 1/2, 1/3, and 1/6. The term PX is one of the price deflators in the model. Note that each equation includes as explanatory variables both the lagged dependent variable and the current and lagged values of r, which implies a fairly complicated lag structure of r on both the long-term rates.

When considered by themselves, equations (5) and (6) are consistent with the expectations theory in the sense that the current value of r and the lagged values of r and of the inflation rate in (5) and (6) are proxying for the expected future values in (4). When considered as part of the overall model, however, equations (5) and (6) are not consistent with the expectations theory *if expectations of the future values of r are rational*. This is because in simulations of the model the predicted values of RA_t, RM_t, r_t, $r_{t+1}, \ldots, r_{t+n-1}$ do not in general satisfy equation (4).

It may help in understanding why Model 1 is not consistent with the existence of rational expectations in the bond market to consider how it can be modified to be consistent. This modification consists of dropping the term-structure equations (5) and (6) from the model and requiring instead that the solution values of r and RA and of r and RM satisfy equation (4).[8] The resulting model, which will be called Model 2, is then consistent with the assumption of rational expectations in the bond market *if* the following hold:

ASSUMPTION 1: *People believe that Model 2 is the true model and know how to solve it.*

ASSUMPTION 2: *People at any one time have the same set of forecasts regarding the future values of the exogenous variables in*

[8]It is easier to say this than it is to program it. Requiring that the solution values satisfy equation (4) means that the solution values of r for periods $t + 1$ and beyond affect the solution values of RA and RM for period t. In this sense future predicted values affect present predicted values, and so the model differs from the typical econometric model, where only present and past values affect present values. In order to solve the model in this case, one must iterate on solution *paths*. For, say, a forty-quarter problem, one first solves the model, given the exogenous variable values, for the forty quarters using guessed values of RA and RM. New values of RA and RM are then computed using equation (4) and the predicted values of r. The model is then resolved for the forty quarters using these new values of RA and RM. New values of RA and RM are then computed from equation (4) using the new predicted values of r, and the model is solved again. There is no guarantee that this process will converge, but for the work in this study it did converge after some damping of some of the solution values. It took an average of about fifteen iterations for this process to converge, which means that the model is about fifteen times more expensive to solve in this case than it is in the regular case. As noted in fn. 2, Poole (p. 478) questioned the computational feasibility of this procedure for large-scale models, but it is in fact not all that expensive. It should also be noted that for the results in this study, values of r beyond the end of the data period were needed. The procedure that was followed to construct these values is discussed in Section III.

the model and the same set of expectations regarding the future values of the error terms.

Given the set of exogenous variable forecasts and the set of expectations of error terms, the solution values of the endogenous variables from the model are also people's expectations of these values.[9] Three of the endogenous variables in the model are r, RA, and RM, and so if the solution values of these variables satisfy equation (4), then people's expectations are consistent with this equation. Model 1, on the other hand, even if Assumptions 1 and 2 above were true of it, would not be consistent with the rational expectations assumption because the solution (i.e., expected) values would not satisfy equation (4). There would still be, in other words, an inconsistency between the assumption of rational expectations and the postulated link in (4) between long rates and expected future short rates.

B. *The Stock-Price Equation*

In a manner analogous to the above analysis of the bond market, in order to consider the policy implications of the assumption of rational expectations in the stock market, some assumption about the determination of stock prices must first be made. The following analysis is based on the theory that the price of a stock is the present discounted value of its expected future returns, although again this analysis could be easily modified to incorporate a slightly different theory. All that is needed for the work below is some link between stock prices and expected future returns.

[9]One subtle point should be noted here. When a non-linear model is solved by setting the error terms equal to their expected values, the solution values of the endogenous variables are not in general equal to their expected values. The proper way to solve non-linear models is by means of stochastic simulation, but because of the expense, this is rarely done. Since no stochastic simulation was done in this study, the predicted values of the one-period rates generated by the model are not exactly equal to their expected values. For simplicity, however, no distinction will be made in the text between predicted and expected values.

Let SP_{t-1} denote the value of corporate stocks held by the household sector at the end of period t − 1, and let ${}_t\Pi^e_{t+i}$ denote the expected value of after-tax cash flow for period t + i, the expectation being conditional on information available as of the beginning of period t. Then according to the above theory:

$$(7)\quad SP_{t-1} = \frac{{}_t\Pi^e_t}{(1+{}_tr^e_t)} + \frac{{}_t\Pi^e_{t+1}}{(1+{}_tr^e_t)(1+{}_tr^e_{t+1})} + \dots + \frac{{}_t\Pi^e_{t+T}}{(1+{}_tr^e_t)(1+{}_tr^e_{t+1})\dots(1+{}_tr^e_{t+T})}$$

where T is large enough to make the last term in (7) negligible. An equation like (7) also holds, of course, for SP_t, with t + 1 replacing t:

$$(8)\quad SP_t = \frac{{}_{t+1}\Pi^e_{t+1}}{(1+{}_{t+1}r^e_{t+1})} + \frac{{}_{t+1}\Pi^e_{t+2}}{(1+{}_{t+1}r^e_{t+1})(1+{}_{t+1}r^e_{t+2})} + \dots + \frac{{}_{t+1}\Pi^e_{t+T+1}}{(1+{}_{t+1}r^e_{t+1})(1+{}_{t+1}r^e_{t+2})\dots(1+{}_{t+1}r^e_{t+T+1})}$$

By definition:

$$(9)\qquad CG_t = SP_t - SP_{t-1}$$

where CG is, as mentioned in Section I, the value of capital gains or losses on corporate stocks held by the household sector.

Since the expected values in equations (7) and (8) are unobserved, some assumption about how expectations are formed must be made in order to implement the above theory. In Model 1 the current change in RA is used as a proxy for changes in the expected future values of r, and a weighted average of the current and past changes in after-tax cash flow is used as a proxy for changes in expected future after-tax cash flow. The estimated equation for CG_t is

$$(10)\quad CG_t = \underset{(2.53)}{13.19} - \underset{(3.19)}{124.3}\Delta RA_t + \underset{(2.12)}{9.824}\left(\frac{1}{2}\Delta\Pi_t + \frac{1}{3}\Delta\Pi_{t-1} + \frac{1}{6}\Delta\Pi_{t-2}\right)$$

$R^2 = 0.212$, $S.E. = 44.02$, $D.W. = 2.33$
Sample Period = 1954I–1976II

The last term in equation (10) is a weighted average of the change in Π for the current and past two quarters, with weights of 1/2, 1/3, and 1/6, where Π is the value of after-tax cash flow of the firm sector and is an endogenous variable in the model.[10]

When considered by itself, equation (10) is consistent with equations (7)–(9) in the sense that ΔRA_t and the weighted average term in (10) are proxying for the changes in expected future interest rates and after-tax cash flow that are implicit in (9). Again, however, when equation (10) is considered as part of the overall model, it is not consistent with equations (7)–(9) if expectations of the future values are rational. This is because in simulations of the model the predicted values of CG_t, $\Pi_t, \ldots, \Pi_{t+T+1}$, $r_t, \ldots, r_{t+T+1}$ do not in general satisfy (7)–(9).

It is possible to drop equation (10) from the model and to require instead that the solution values of CG, Π, and r satisfy equations (7)–(9).[11] If this is done for Model 2, then the resulting model, which will be called Model 3, is consistent with the assumption of rational expectations in both the bond and stock markets, provided that it is also assumed that Assumptions 1 and 2 above are true for Model 3. Note for Model 3 that because r is used as the discount rate in (7) and (8), the expected return on stocks is the same as the expected return on bonds. In other words, there are no arbitrage opportunities in Model 3 between bonds and stocks, just as there are no arbitrage opportunities in either Model 2 or 3 between bonds of different maturities.

III. The Experiments

In order to examine the sensitivity of policy effects to the assumption of rational expectations in the bond and stock markets, two basic experiments were performed for each of the three models: the first is a fiscal policy action and the second a monetary policy action. For the first experiment, the real value of goods purchased by the government (XG) was permanently increased by $1.25 billion beginning in 1971I, a quarter that is at or near the bottom of a contraction. The behavior of the Fed was assumed to be endogenous for this experiment: the equation explaining Fed behavior, equation (3), was added to the model, and the amount of government securities outstanding VBG was taken to be endogenous. For the second experiment, VBG was permanently decreased by $1.25 billion beginning in 1971I. In this case the behavior of the Fed is obviously not endogenous, so equation (3) is not included in the model.

A standard procedure in performing experiments of this type, which was followed here, is first to add to the stochastic equations of the model the residuals obtained in the process of estimating the equations. Doing this means that when the model is simulated using the actual values of all exogenous variables, the predicted values of all endogenous variables are equal to their actual values. In other words, a perfect tracking solution is obtained. These same residuals are then used for all the experiments; they are treated in effect like exogenous variables. This procedure allows the predicted values of the endogenous variables obtained from changing one or more exogenous variables to be compared directly to their actual values in examining the effects of the change. Also, since multipliers in non-linear models are a function of initial conditions and of values of the exogenous variables and error terms, this procedure provides a natural base from which to perform the experiments, namely the actual data.

The experiments for Model 1 are straightforward to perform. The estimated residuals are first added to the model, and then the model is simulated for the exogenous variable changes. As noted above, the simulations began in 1971I. They were dynamic, twelve-quarter simulations. The results for the first experiment are presented in Table 1, and the

[10] In terms of the notation in my 1976 book, $\Pi = CF - TAXF$, where CF and $TAXF$ are both endogenous variables.

[11] In Model 3, unlike in Model 2, future predicted values of Π (as well as r) affect present predicted values. The solution procedure outlined in fn. 8 for Model 2 can, however, with obvious modifications, also be applied to Model 3.

results for the second experiment are presented in Table 2.[12]

The experiments for Models 2 and 3 require that some further assumptions be made. First, some assumption must be made about people's expectations of the future values of the exogenous variables and error terms. For present purposes these expectations were assumed to be perfect. In other words, people were assumed to know the actual future values of all the exogenous variables and all the estimated residuals. As discussed below, this assumption in the present context is actually not very restrictive. Second, some assumption about n in equation (4) and T in equations (7) and (8) has to be made. For the experiments, n was assumed to be 32 for both RA and RM, and T was assumed to be 80. In other words, both RA and RM were assumed to be rates on eight-year securities, and the horizon for determining stock prices was assumed to be twenty years. Third, some assumption has to be made about the predicted (i.e., expected) values of r and Π beyond the end of the period for which there are data on the exogenous variables and estimated residuals. One possibility in this case would be to project the exogenous variables and error terms as far into the future as needed to make the end-point effects have a negligible influence on the period of interest. In this study, however, a somewhat simpler procedure was followed. The expected values of r beyond the end of the data (1976II) were assumed to be equal to the average of the last eight expected values within the data period (i.e., to the average of the expected values of r for the 1974III–1976II period). Similarly, the expected values of Π beyond the end of the data were assumed to be equal to the average of the last eight expected values within the data period. As will be discussed in Section V, the results of the experiments do not appear to be very sensitive to this assumption.

In order to make the results for Models 2 and 3 comparable to those for Model 1, perfect tracking solutions for the two models must first be obtained. This can be done as follows. First, given a value of n, given the actual data on r, and given the above assumption about the values of r beyond the end of the data, equation (4) can be used to compute predicted values of R, say R^*. The difference between RA_t and R_t^* is the estimated residual in predicting RA_t for period t. If this residual is added to equation (4) in the appropriate way, a perfect fit for RA_t is obtained. Likewise, the difference between RM_t and R_t^* is the estimated residual in predicting RM_t for period t, and if this residual is added to equation (4) in the appropriate way, a perfect fit for RM_t is also obtained. Similarly, given a value of T, given the actual data on r and Π, and given the above assumptions about the values of r and Π beyond the end of the data, equations (7)–(9) can be used to compute predicted values of CG, say CG^*. The difference between CG_t and CG_t^* is the estimated residual in predicting CG_t for period t, and if this residual is added to equation (9), a perfect fit for CG_t is obtained. Perfect tracking solutions for Models 2 and 3 can thus be obtained by using these estimated residuals for equations (4) and (9) along with the estimated residuals from Model 1 for the other equations.

Given the above assumptions, the experiments for Models 2 and 3 can now be described. The estimated residuals are first added to each model, and then the model is simulated for the exogenous variable changes. As was the case for Model 1, the simulations began in 1971I and were dynamic. The simulations were allowed to run to the end of the data (1976II), at which point the assumptions about r and Π beyond the end of the data came into play. Results for the first twelve quarters of the simulation period are presented for each model in Tables 1 and 2.

It should be noted that this simulation procedure implicitly assumes that the changes in XG and VBG that began in 1971I were unanticipated changes. If instead it were assumed that the government announced in, say, 1969I that it would make the changes in XG or VBG beginning in 1971I, then the simulations for Models 2 and 3 would have had to begin in 1969I. In other words, in Models 2 and 3 people would have begun in 1969I adjusting to the announced future

[12] The results in Table 1 for Model 1 are the same as the results presented in my 1978a paper, Table 1, for the endogenous Fed case.

policy changes. Some results are reported in Section V for the case in which the anticipation of the changes precedes the actual changes, but for the basic results in Section IV the changes are assumed to have been unanticipated.

One further point about the solution for Model 3 should also be noted. If, say, the first quarter of the simulation period is quarter t, then for Models 1 and 2, SP_{t-1}, the value of stocks at the beginning of quarter t, is predetermined. For Model 3, however, SP_{t-1} is endogenous. It is determined according to equation (7), where the expected future values of Π and r in the equation are the values predicted by the model. This means that CG_{t-1} is also endogenous for Model 3, since it equals $SP_{t-1} - SP_{t-2}$. Consequently, values of CG are presented in Tables 1 and 2 for quarter $t - 1$ as well as for the other twelve quarters for Model 3.

Before proceeding to the discussion of the basic results, it can now be seen why the assumption that people know the actual future values of all the exogenous variables and all the estimated residuals is not very restrictive in the present context. If instead some different set of values were used, based on extrapolations for the exogenous variables and zeros for the residuals, say, and if this same set were used for all three models, it is unlikely that the comparisons across models would be much different. In this case one would take as the estimates of the effects of the policy change for each model the differences between the predicted values of the endogenous variables before and after the change. The predicted values of the endogenous variables before the change would no longer be the actual values. Because the models are non-linear, the actual numbers in Tables 1 and 2 would be different in this case, but these differences are likely to be fairly similar for each of the three models. Therefore, little is likely to be lost in the present context by running the experiments off of the perfect tracking solution.[13]

IV. The Basic Results

The effects on seven variables are presented in Tables 1 and 2. Each number in the tables is the difference between the predicted value of the variable for the quarter and the actual value; Y is the key output variable in the models and PF is the key price variable.

Consider the results in Table 1 first. The increase in XG led to an increase in output in all three models. The sum of the output increases over the twelve quarters was 16.81 for Model 1, 9.59 for Model 2, and 10.27 for Model 3. The increases in RA and RM for the first four quarters were much smaller for Model 1 than they were for the other two models. This is, of course, as expected. In Models 2 and 3 people expect that the Fed is going to respond to the fiscal policy stimulus by increasing r (i.e., people know equation (3)), and RA and RM adjust immediately to these expectations. In Model 1, on the other hand, RA and RM adjust only to the current and lagged increases in r. Higher values of RA and RM have, other things being equal, a contractionary effect on the economy, and this is the primary reason for the lower values of Y for Models 2 and 3 compared to those for Model 1. The bond rate RA also has, other things being equal, a positive effect on PF, and this is the main reason for the higher values of PF for Models 2 and 3 compared to those for Model 1.

---

[13]Although the assumption that people know the actual future values of all exogenous variables and all estimated residuals does not seem restrictive for present purposes, it would clearly not be reasonable to use it in any test of the assumption of rational expectations. One possible test that could be performed using Model 3, which is beyond the scope of this paper, is the following: 1) Choose for each quarter a set of future values of the exogenous variables and error terms that one believes were expected at the time. (In most cases the future values of the error terms would be zero.) 2) Using Model 3, compute for each quarter the predicted values of RM, RA, and CG. The predictions of these variables for each quarter would be based on different initial conditions and a different set of future values of the exogenous variables. 3) Compare the accuracy of these predictions to the accuracy of predictions from other models. The joint hypothesis that would be examined or tested by this procedure is that (a) people know Model 3 and believe it to be true, including equations (4) and (7)–(9) that link expected future values to current values, (b) the chosen exogenous variable values and error terms correctly reflect the expectations at the time, and (c) expectations with respect to the future values of r and Π are rational.

TABLE 1—FISCAL POLICY RESULTS: DIFFERENCE BETWEEN THE PREDICTED VALUE AFTER THE CHANGE AND THE ACTUAL VALUE[a]

VARIABLES	QUARTERS t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	SUM 1-12
r (bill rate)														
1		0.10	0.20	0.26	0.30	0.31	0.31	0.30	0.28	0.27	0.25	0.24	0.24	
2		0.08	0.15	0.18	0.19	0.19	0.19	0.18	0.18	0.17	0.16	0.16	0.16	
3		0.08	0.15	0.18	0.20	0.20	0.20	0.20	0.19	0.19	0.18	0.18	0.18	
RA (bond rate)														
1		0.03	0.06	0.07	0.11	0.14	0.15	0.16	0.17	0.18	0.18	0.18	0.19	
2		0.21	0.21	0.22	0.22	0.22	0.22	0.22	0.22	0.23	0.23	0.23	0.23	
3		0.23	0.24	0.24	0.25	0.25	0.25	0.25	0.26	0.26	0.26	0.27	0.27	
RM (mortgage rate)														
1		0.01	0.03	0.05	0.07	0.10	0.13	0.14	0.15	0.15	0.16	0.16	0.17	
2		0.21	0.21	0.22	0.22	0.22	0.22	0.22	0.22	0.23	0.23	0.23	0.23	
3		0.23	0.24	0.24	0.25	0.25	0.25	0.25	0.26	0.26	0.26	0.27	0.27	
Y (real output)														
1		1.24	1.77	2.04	2.12	1.95	1.72	1.48	1.25	1.03	0.84	0.71	0.66	16.81
2		0.98	1.12	1.11	1.08	0.97	0.86	0.77	0.68	0.58	0.49	0.46	0.49	9.59
3		0.93	1.10	1.12	1.12	1.03	0.94	0.86	0.77	0.67	0.58	0.56	0.59	10.27
100·PF (price deflator)														
1		0.02	0.05	0.09	0.14	0.19	0.25	0.29	0.33	0.37	0.41	0.45	0.48	
2		0.07	0.13	0.18	0.23	0.27	0.31	0.34	0.37	0.39	0.42	0.45	0.47	
3		0.08	0.15	0.20	0.25	0.30	0.34	0.38	0.41	0.44	0.47	0.50	0.53	
CG (capital gains or losses on stocks)														
1	0.00	-2.14	-3.34	-2.55	-6.19	-5.85	-1.74	-1.47	-0.48	-0.37	0.03	0.09	-0.77	-24.78
2	0.00	-24.40	-1.18	-1.04	-0.82	-0.35	-0.06	0.02	0.06	-0.07	-0.15	-0.14	-0.09	-28.22
3	-3.13	-0.34	-0.05	-0.01	-0.00	0.04	0.00	-0.03	-0.06	-0.08	-0.10	-0.15	-0.16	-4.07
VBG (amount of government securities outstanding)														
1		0.66	1.07	1.34	1.61	1.90	2.29	2.79	3.41	4.12	4.93	5.82	6.75	
2		0.67	1.29	1.90	2.58	3.29	4.06	4.87	5.72	6.64	7.61	8.62	9.66	
3		0.68	1.28	1.85	2.46	3.09	3.76	4.46	5.20	6.00	6.85	7.72	8.61	

Notes: Units of variables are: percentage points at an annual rate for r, RA, and RM; billions of 1972 dollars at a quarterly rate for XG and Y; 1972 = 1.0 for PF; billions of current dollars at a quarterly rate for CG; billions of current dollars for VBG.

[a]Effects of a permanent increase in XG of $1.25 billion beginning in quarter t (t = 1971). 1 = Model 1 (original version); 2 = Model 2 (rational expectations in bond market); 3 = Model 3 (rational expectations in bond and stock markets).

It is interesting to note that the values of r are higher for Model 1 than they are for Models 2 and 3. Since the economy is less expansionary for Models 2 and 3 than it is for Model 1, due to the more rapid response of RA and RM, the Fed raises r less in these two cases than it does in the Model 1 case.

The economy is slightly more expansionary for Model 3 than it is for Model 2, and this is easy to explain. For Model 2 there was a large capital loss on stocks in quarter t because of the increase in RA. (Remember that the estimated equation for CG_t, equation (10), is still part of Model 2.) For Model 3, however, the negative effects of the higher expected future values of r on the value of stocks were almost completely offset by the positive effects of higher expected future values of after-tax cash flow caused by the increase in economic activity. The capital loss incurred at the beginning of quarter t was 3.13 for Model 3, compared to the capital loss in quarter t of 24.40 for Model 2. Since capital losses have a negative effect on the economy through the wealth effect on the household sector, the economy was somewhat more expansionary

TABLE 2—MONETARY POLICY RESULTS: DIFFERENCE BETWEEN THE PREDICTED VALUE AFTER THE CHANGE AND THE ACTUAL VALUE[a]

VARIABLES		QUARTERS												SUM
	t-1	t	t+1	t+2	t+3	t+4	t+5	t+6	t+7	t+8	t+9	t+10	t+11	1-12
r (bill rate)														
1		-1.74	0.37	1.41	1.43	0.78	-0.05	-0.55	-0.61	-0.40	-0.13	0.12	0.25	
2		-1.35	-0.50	0.06	0.34	0.35	0.27	0.18	0.10	0.05	0.04	0.04	0.05	
3		-1.35	-0.52	0.04	0.33	0.36	0.30	0.22	0.13	0.08	0.06	0.05	0.05	
RA (bond rate)														
1		-0.73	0.21	0.12	0.19	0.16	-0.01	-0.07	-0.06	-0.03	-0.02	-0.01	0.01	
2		-0.02	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	
3		-0.02	0.04	0.06	0.05	0.04	0.03	0.01	0.01	0.01	0.00	0.00	0.00	
RM (mortgage rate)														
1		-0.14	-0.48	0.21	0.08	0.14	0.11	-0.00	-0.03	-0.02	-0.01	-0.02	-0.01	
2		-0.02	0.04	0.05	0.05	0.04	0.03	0.02	0.01	0.01	0.01	0.01	0.01	
3		-0.02	0.04	0.06	0.05	0.04	0.03	0.01	0.01	0.01	0.00	0.00	0.00	
Y (real output)														
1		0.77	1.98	1.39	0.44	-0.51	-0.85	-0.62	-0.19	0.14	0.27	0.26	0.15	3.23
2		0.39	0.59	0.54	0.35	0.13	0.00	-0.06	-0.07	-0.06	-0.04	-0.02	-0.02	1.73
3		0.39	0.57	0.54	0.36	0.15	0.02	-0.04	-0.06	-0.05	-0.03	-0.02	-0.01	1.82
100·PF (price deflator)														
1		-0.24	-0.10	-0.02	0.06	0.11	0.08	0.03	0.00	-0.00	0.00	0.01	0.02	
2		-0.00	0.02	0.04	0.05	0.06	0.06	0.06	0.05	0.05	0.05	0.04	0.04	
3		-0.00	0.02	0.04	0.05	0.06	0.06	0.06	0.05	0.05	0.04	0.04	0.04	
CG (capital gains or losses on stocks)														
1	0.00	92.54	-117.66	8.89	-11.96	2.57	23.66	10.47	0.13	-3.09	-2.37	-2.63	-2.70	-2.15
2	0.00	2.76	-7.30	-2.35	-0.44	1.19	1.95	1.53	0.93	0.45	0.22	0.07	0.04	-0.95
3	-2.00	-0.78	-0.21	0.15	0.34	0.33	0.21	0.11	0.04	0.02	0.02	0.03	0.04	-1.70
VBG (amount of government securities outstanding)														
1		-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	
2		-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	
3		-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	-1.25	

Notes: See Table 1.

[a]Effects of a permanent decrease in *VBG* of $1.25 billion beginning in quarter t (t = 1971I). 1 = Model 1 (original version); 2 = Model 2 (rational expectations in bond market), 3 = Model 3 (rational expectations in bond and stock markets.

for Model 3 than it was for Model 2. This difference is, however, much smaller than the difference between the results for Models 1 and 2. In other words, adding rational expectations in the bond market to the original version of the model makes more of a difference than does the further addition of rational expectations in the stock market. In this sense, wealth effects in the model are less important than interest rate effects.

Consider now the monetary policy experiment in Table 2. The experiment itself is not as expansionary as the experiment in Table 1, but the comparison across models in Table 2 is similar to that in Table 1. The economy is more expansionary for Model 3 than it is for Model 2. For Model 1 the decrease in *VBG* led to a large decrease in *r* in quarter t and then a bounce back again in quarter t + 1. The rate *RA* was affected in a similar way, which resulted in a large capital gain in quarter t and a large capital loss in quarter t + 1. For Models 2 and 3, on the other hand, *RA* and *RM* were much less affected by the large initial change in *r*. Rates *RA* and *RM* in fact changed very little for Models 2 and 3 because the long-run effect of the change in *VBG* on *r* was fairly small. In other words, people expect in Models 2 and 3 that the large initial drop in *r* is temporary, and so *RA* and

RM do not respond very much to this event. In Model 1 people do not expect this, and so the initial changes in *RA* and *RM* are much larger.

Comparing the results for Models 2 and 3 in Table 2, it can be seen that after the first four quarters Model 2 has a cumulative capital loss of 7.33 compared to a cumulative capital loss of only 2.50 for Model 3. This is the main reason for the slightly more expansionary economy for Model 3. The difference between the results for the two models is, however, quite small, and the cumulative capital loss over twelve quarters is in fact slightly larger for Model 3 than it is for Model 2.

In summary, then, the results in Tables 1 and 2 indicate that the long-run effects of fiscal policy and monetary policy actions on real output are a little over half as large if there are rational expectations in the bond and stock markets than if there are not. For the fiscal policy experiment, the sum of the output increases over twelve quarters for Model 3 is 61.1 percent of that for Model 1 (10.27/16.81). The corresponding figure for the monetary policy experiment is 56.3 percent (1.82/3.23). The results also indicate that the addition of rational expectations in the bond market to the model is quantitatively more important than is the addition of rational expectations in the stock market.

V. Further Results

The results of three other experiments are reported in this section. The first is the same as the first experiment in Section IV except that the behavior of the Fed is assumed to be different. Instead of assuming that the Fed behaves according to equation (3), it was assumed that the Fed behaves by keeping r unchanged each period from its historic value. In this case the behavior of the Fed is exogenous in the sense that its behavior with respect to the value of r each period is not a function of any endogenous or lagged endogenous variables in the model. However, *VBG* is still endogenous even though equation (3) is dropped from the model, since r is now exogenous.

The results for Models 1 and 2 were nearly identical for this experiment. The sum of the changes in r over the first twelve quarters was 27.73 for Model 1 and 28.07 for Model 2. For Model 2, *RA* and *RM* were completely unchanged, as is obvious from equation (4), but for Model 1 they were slightly higher as a result of the inflation term in equations (5) and (6). The slightly higher values of *RA* and *RM* thus led to a slightly smaller output increase in Model 1 than in Model 2. This difference is, however, almost negligible. The main point of this example is that if the Fed keeps r unchanged, then the policy implications of Models 1 and 2 are quite similar. In Model 2 people expect that the short rates will not change in response to the fiscal policy stimulus, and so there is no change in the long rates. In Model 1 people have no explicit expectations of this sort, but a property of the estimated term-structure equations is that long rates do not change much if short rates do not change.

The second experiment is the same as the first experiment in Section IV except that the starting quarter for the change in *XG* was taken to be 1958I rather than 1971I. As was the case for the results in Section IV, the simulations were run to the end of the data (1976II) for Models 2 and 3, at which point the assumptions about r and Π beyond the end of the data came into play. For this longer period the results for Models 2 and 3 for, say, the first twelve quarters should be less sensitive to the end-point assumptions. It turned out, however, that these results were quite similar to the results for the shorter period presented in Section IV. For the experiment that began in 1958I, the sum of the changes in Y over the first twelve quarters was 16.27 for Model 1 and 8.04 for Model 3. The Model 3 response was thus 49.4 percent of the Model 1 response (8.04/16.27), which is only slightly lower than the figure of 61.1 percent for the first experiment in Section IV. The other results were also similar between the two experiments. It thus appears that the use in Section IV of only ten quarters beyond the basic twelve-quarter prediction period for Models 2 and 3 is enough to capture most of the effects of the future predicted values on the present predicted values.

The third and final experiment is the same

as the first experiment in Section IV except that it was assumed that the government announced the fiscal policy action (to begin in 1971I) in 1958I. In other words, the policy change was assumed to be announced thirteen years before it was actually made. The starting quarter for this experiment was 1958I. For Model 1 the results for the 1971I–1973IV period are exactly as reported in Table 1, since in this model future changes in exogenous variables do not affect current predicted values. For Models 2 and 3, however, the results are different. For Model 3, for example, there was a cumulative output loss between 1958I and 1970IV of 5.70. This output loss was due to the fact that people expected the Fed to raise the bill rate in response to the fiscal policy stimulus, and these expectations got reflected in the values of the long rates before 1971I. The higher long rates then had a contractionary effect on the economy prior to the actual change in XG. After the change in XG in 1971I there was a cumulative output gain during the next twelve quarters (1971I–1973IV) of 10.50. The net cumulative gain through 1973IV was thus 4.80, compared to the 10.27 figure in Table 1 for the unanticipated change. The cumulative output gain that occurs when the policy change is anticipated is thus slightly less than half of the gain that occurs when the policy change is unanticipated.

VI. Summary and Conclusion

This study has demonstrated that it is feasible to analyze a large-scale macro-econometric model with rational expectations in the bond and stock markets. The primary purpose of this paper has been to present some quantitative estimates of the sensitivity of monetary and fiscal policy effects to the assumption of rational expectations in the bond and stock markets. With some qualifications, the results indicate that unanticipated policy actions are about half as effective and anticipated policy actions are about one-fourth as effective (with respect to real output changes) when there are rational expectations in these markets than when there are not. The results also indicate that the existence of rational expectations in the bond market is quantitatively more important than is the existence of rational expectations in the stock market.

The results of this paper must, however, be interpreted with the usual degree of caution. First, they are dependent on a particular model, and there is at least a small probability that this model is not a good representation of the economy. Second, it is well known that multipliers in non-linear models depend on initial conditions, and the basic starting point for the results in this paper was a quarter that was at or near the bottom of a contraction. Clearly, different results would have been obtained had, say, the starting point been the top of an expansion. Third, it is also well known that fiscal policy effects depend on what one assumes about Fed behavior, and, as reported in Section V, quite different results can be obtained if it is assumed that the Fed behaves differently than is estimated by equation (3).

Given these caveats, it is hoped that this study has made some progress in determining the quantitative importance of the rational expectations assumption with respect to the bond and stock markets. It appears from the present results that this assumption is of considerable quantitative significance.

REFERENCES

P. A. Anderson, "'Rational' Forecasts from 'Nonrational' Models," work. paper no. 61, Fed. Reserve Bank Minneapolis, rev. Jan. 1977.

R. J. Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

Ray C. Fair, "The Estimation of Simultaneous Equation Models with Lagged Endogenous Variables and First Order Serially Correlated Errors," *Econometrica*, May 1970, *38*, 507–16.

———, *A Model of Macroeconomic Activity*, Vol. I: *The Theoretical Model*, Cambridge, Mass. 1974; Vol. II: *The Empirical Model*, Cambridge, Mass. 1976.

———, (1978a) "The Sensitivity of Fiscal-Policy Effects to Assumptions about the Behavior of the Federal Reserve," *Econo-*

metrica, Sept. 1978, *46*, 1165–79.

———, (1978b) "A Criticism of One Class of Macroeconomic Models with Rational Expectations," *J. Money, Credit, Banking*, Nov. 1978, *10*, 411–17.

R. E. Lucas, Jr., "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

F. Modigliani and R. J. Shiller, "Inflation, Rational Expectations and the Term Structure of Interest Rates," *Economica*, Feb. 1973, *40*, 12–43.

W. Poole, "Rational Expectations in the Macro Model," *Brookings Papers*, Washington 1976, *2*, 463–505.

T. J. Sargent, "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment," *Brookings Papers*, Washington 1973, *2*, 429–80.

———, "A Classical Macroeconometric Model for the United States," *J. Polit. Econ.*, Apr. 1976, *84*, 207–37.

——— **and N. Wallace**, "'Rational' Expectations, the Optimal Monetary Instrument, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

The Narrowing of Black-White Wage Differentials is Illusory

By EDWARD LAZEAR*

Wage differentials have been a topic of interest to economists for many years. Analyses have been conducted with respect to skill and industry differentials (see Melvin Reder, 1955, 1960; Sherwin Rosen, 1970; P. G. Keat), union differentials (see H. Gregg Lewis; Rosen, 1969; Orley Ashenfelter and George Johnson), and more recently black-white differentials (see James Smith and Finis Welch; Welch; Richard Freeman 1975, 1976). Virtually all of these studies are cross sectional in nature and, as such, examine wage differences as viewed at a point in time. More recent labor economics literature has focused on the importance of wealth rather than income, and more specifically on the conscious decisions that individuals make which alter the nature of their earnings stream.[1] Thus, we do not view two individuals with the same current measured incomes as having the same real incomes if one is on a job track that will yield him $50 thousand per year at age 30 and the other is on a profile that yields him $10 thousand per year at that same age. Individuals are clearly not indifferent between varying rates of wage growth. In fact, one may calculate the value of differential wage growth by determining the amount by which wealth is altered as the result of the differential. If jobs offer differing wage growth opportunities, the compensation for those jobs should be measured in a way that includes this unobserved but anticipated compensation.

Recent findings (see Freeman, 1975; Smith and Welch; Welch) have revealed that there has been substantial narrowing of the black-white wage differential during the past ten to fifteen years and that this narrowing has been most pronounced among young workers. In the data used for this analysis, one finds that the differential has completely disappeared by 1972. This result is difficult to accept at face value. In this paper, it is suggested that much of this narrowing may be due to a wage measurement problem. It will be argued that the narrowing of the current wage differential has been coupled with a widening of job-experience-induced rates of wage growth. This implies that there has not been as great a narrowing in the black-white differential as it appears from looking at observed wages. Instead, blacks in more recent cohorts have experienced a relative substitution of current wages for future wages or earnings power. But this differential in total compensation is severely overstated by differences in pecuniary wages. It appears that much of what employers have given nonwhites in current wages has been recaptured by a reduction in on-the-job training (*OJT*) provided.

One explanation for this change in the structure of employee compensation is that the move is a result of government affirmative action programs. Suppose, for example, that governmental legislation made it increasingly difficult to discriminate in the form of pecuniary wages. We expect employers to raise the observed wage for blacks while at the same time decreasing the unobserved (or more difficult to observe) *OJT* component of earnings, thereby keeping the true differential constant. If government programs concentrate on entry level job discrimination, and if differential *OJT* does not show up until later, examining the pecuniary wage rate for young

*Associate professor, University of Chicago and research associate, National Bureau of Economic Research. I wish to thank Charles Brown, Richard Freeman, Zvi Griliches, Robert Hall, James Heckman, Victoria Lazear, Tom MaCurdy, Robert Michael, Melvin Reder, and Lester Telser for helpful comments Craig Hakkio performed the computation. Financial support from the National Institute of Mental Health and the Rockefeller Foundation is gratefully acknowledged.

[1]This is exemplified by Jacob Mincer (1960); Yoram Ben-Porath; William Haley; James Heckman; Lee Lillard.

workers may greatly distort our view of the true wealth differential. Indeed, almost any conceivable model of optimal law evasion would suggest that employers would respond to legislation requiring equal pecuniary wage rates for nonwhites and whites at least in part by altering nonpecuniary aspects of the wage. The *OJT* is, at least for young workers, a large part of the nonpecuniary component.[2] Under this scenario, the observed wage differential narrows, but the narrowing, the result of government pressure to reduce current wage differentials, disguises the constancy of the true wage differential which includes compensation of all forms.[3]

This paper employs a method (devised in my 1976 paper) to estimate the unobserved component of wages. The size of this component is calculated for nonwhites and whites separately and then compared. Since, as it turns out, the component is larger for whites than nonwhites, observed wage differentials understate true differentials. More important is that comparison of the period 1966–69 with 1972–74 reveals that this unobserved differential increased substantially over time. The results of this study suggest that although the pecuniary nonwhite-white differential has narrowed substantially between 1966 and 1974 for young men, the *OJT* differential has increased by almost the same amount. This implies that in real wealth terms there has not been any narrowing of the white-nonwhite differential at all. This will become more apparent in later years as those nonwhites who were hired into skilled jobs today fail to be promoted or to obtain higher paying jobs elsewhere at the same rate as their white counterparts. Thus nonwhite age-earnings profiles should diverge more from white profiles than they did in the past.

I. A Model

Consider an individual who has an observed wage rate $f(t)$ at time t. Suppose that at some point in time, t_0, the individual is offered two alternatives: he may have $f(t_0)$ and a path in the future such that $\partial f(t_0)/\partial t$ is positive, or he may have $f(t_0)$ and a path such that $\partial f(t_0)/\partial t$ is zero. He clearly prefers the first situation. How much would he be willing to pay to obtain the first over the second? If the wage increase is permanent, then the individual's wages are higher by $\partial f(t_0)/\partial t$ for every period from t_0 until retirement. The value of this wage increase in t_0 value terms is then

$$[\partial f(t_0)/\partial t] \int_0^{T-t_0} e^{-r\tau} d\tau$$

where T is the date of retirement. Thus "true" earnings at t_0 should include the value of the wage growth in addition to the observed component, $f(t_0)$.

There is a wrinkle. The individual often pays for the wage growth by accepting a wage lower than he could have earned had he settled for zero wage growth. This is because when he is learning on the job, he is producing less than he would have produced had he devoted all his time and effort to commodity production. (For example, given an hour at work, he may spend fifteen minutes of it learning to use a machine. Thus, the employer views this as forty-five minutes and his hourly

[2]Stated otherwise, the argument is as follows: Before governmental intervention, the nonwhite worker optimally chose to produce W_0 per hour in commodities and H_0 in on-the-job training. His earnings were, therefore, $W_0 + H_0$ per hour on the assumption that the worker is paid his marginal product. (Actually, this assumption is unnecessary.) Suppose, now, that the government requires the employer to pay the nonwhite $W_1 > W_0$. The only way this wage will be an equilibrium wage is if the worker produces W_1 per hour. In order to do this, he must transfer $W_1 - W_0$ worth of effort per hour away from the production of *OJT* to that of actual commodities. No other firm can offer him the (W_0, H_0) payment combination since they are subject to the same governmental restriction. Thus, pecuniary wages will rise to at least W_1 for nonwhite workers and human capital production will fall at least to $H_1 = H_0 - (W_1 - W_0)$. I say "at least" since these changes may alter labor force participation and affect other variables which in turn may increase $H_1 - H_0$ to an amount greater than $W_1 - W_0$ as implied by the equilibrium constraint.

[3]Other explanations for this shift in the compensation package seem unlikely. It is difficult to imagine that changes come about as the result of voluntary choice on the part of nonwhites which so dramatically reduce the amount of *OJT* they choose to purchase. This would require dramatic and pessimistic changes in expectations about future labor force participation or about the returns to *OJT* which seem unreasonable. Either way, however, the point remains that the variable relevant for analysis is the total wage, which includes the value of wage growth, rather than the observed wage.

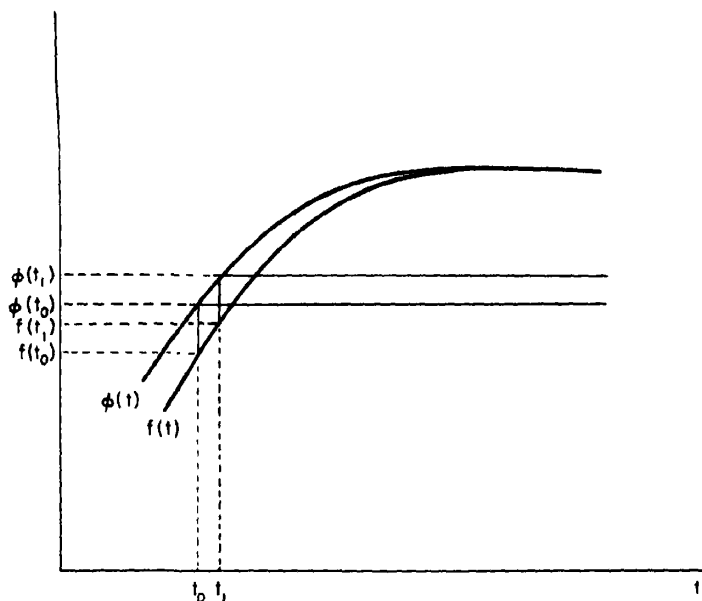



FIGURE 1

wage rate reflects only that amount of production.)

Define, therefore, potential earnings as the amount that the individual would receive at a given point in time if he had spent none of his energy on wage growth production and denote this $\phi(t)$. Note that $\phi(t_0)$ depends upon the stock of human capital that the individual carries with him into period t_0. Take this stock as given at time t_0. The cost of the investment in wage growth at time t is therefore $C(t) \equiv \phi(t) - f(t)$. This is shown on Figure 1. If the individual were to spend all of his time producing at t_0, he would earn $\phi(t_0)$. In this case, he would enjoy no wage growth and his potential and observed earnings would remain $\phi(t_0)$. If instead, he produces only $f(t_0)$ of output and spends $\phi(t_0) - f(t_0)$ on training, his potential earnings will be higher at t_1 than they are at t_0. This higher earnings power is enjoyed until retirement. Its value at t_0 is then

$$H(t) \equiv \phi'(t_0) \int_0^{T-t_0} e^{-r\tau}\, d\tau \tag{1}$$

$H(t_0)$ is then the value of wage growth at time t_0 since an individual would be indifferent between an earnings stream which paid $\phi(t_0) + H(t_0)$ at time t_0 and $\phi(t_0)$ afterward, and one which paid $\phi(t_0)$ at time t_0 and $\phi(t_1)$ afterward. What the individual actually receives is

$$F(t_0) \equiv f(t_0) + \phi'(t_0) \int_0^{T-t_0} e^{-r\tau}\, d\tau$$
$$\equiv f(t_0) + H(t_0)$$

since he pays $C(t_0)$ in order to buy this wage growth.

It is useful to note that in general $C(t_0)$ does not equal $H(t_0)$. It is always the case that $C(t_0) \leq H(t_0)$. The difference between the two is the profit on the investment made at t_0. The individual will never purchase wage growth that is unprofitable.[4] Stated otherwise,

$$f(t_0) + C(t_0) \leq f(t_0) + H(t_0)$$

or

$$\phi(t_0) \leq F(t_0)$$

Investment in human capital can yield inframarginal profits. The individual pushes the rate of investment in any period to the point

[4]See Rosen (1973) or my 1975 paper for a complete discussion of this point.

where the cost of increasing that rate exactly equals the returns. On the last increment of rate increase no profits are earned, but they are earned on all inframarginal increases. Thus, lifetime wealth is increased by

$$\int_0^T [H(t) - C(t)]e^{-rt}dt$$

as the result of the investment in *OJT*. It is $f(t) + H(t)$ that is received by the individual, however, and this is the amount that constitutes real earnings.

If one could observe $\phi(t)$, equation (1) would allow $H(t)$, and therefore $f(t) + H(t)$, to be determined. But only $f(t)$ can be observed. However, although $\phi(t) > f(t)$, under certain circumstances $\phi'(t) \simeq f'(t)$. Since $f'(t)$ can be observed, true earnings could be ascertained under these conditions. Above, $C(t)$ was defined to be $\phi(t) - f(t)$. Thus

$$(2) \qquad C'(t) = \phi'(t) - f'(t)$$

$$\text{or} \qquad f'(t) = \phi'(t) - C'(t)$$

so that if $C'(t)$ equals zero, $f'(t) = \phi'(t)$. If investment in *OJT* is approximately constant over time, then $f'(t)$ will approximate $\phi'(t)$. Theory, however, tells us that $C(t)$ will not be a constant, but will eventually become zero. This implies that $C'(t)$ is negative at least in some range and so $f'(t)$ overstates $\phi'(t)$ for those values of t. Although $C'(t)$ will be negative, one can assign a lower bound and thereby ascertain the maximum bias in $f(t)$.

Define t^* as that time corresponding to the peak of the age-earnings profile. By definition of peak, no net investment is occurring here since earnings neither rise, nor fall. Thus, $\phi(t^*) = f(t^*)$ or $C(t^*) = 0$. Further, we know that at any time $t < t^*$, $C(t) \leq f(t^*) - f(t)$. To the extent that either positive profits exist or that investment does not all occur at t, $C(t)$ will be less than this upper bound. (See Figure 1.) It is now necessary to assume that $C''(t) \leq 0$. (Note that this assumption is consistent with the generally observed concavity of the age-earnings profile.) Since $C''(t) \leq 0$, the average slope of $C(t)$ between t and t^* will be at least as steep as (equal to or more negative than) the slope at t or

$$C'(t) \geq \frac{C(t^*) - C(t)}{t^* - t}$$

Since $C(t^*) = 0$, this implies

$$C'(t) \geq \frac{-C(t)}{t^* - t}$$

But from above

$$C(t) \leq f(t^*) - f(t)$$

$$\text{or} \qquad -C(t) \geq f(t) - f(t^*)$$

Thus,

$$(3) \qquad C'(t) \geq \frac{f(t) - f(t^*)}{t^* - t}$$

The terms $f(t)$ and $f(t^*)$ are observable so that a lower bound to $C'(t)$ can be obtained.[5]

Using $f'(t)$ for $\phi'(t)$,

$$(4) \quad F(t) = f(t) + H(t)$$

$$= f(t) + f'(t) \int_0^{T-t} e^{-r\tau} d\tau$$

Now the point addressed in the introduction can be discussed more rigorously. Cross-sectional analyses of wage differentials only consider differences in the observed values of $f(t)$ at some point in time. Yet equation (4) reveals that this tells only part of the story. A true measure of the wage differential is $F_W(t) - F_{NW}(t)$ where *NW* and *W* refer to nonwhites and whites, respectively. Since it is likely[6] that $H_W(t) > H_{NW}(t)$, the wage differential will be understated by $f_W(t) - f_{NW}(t)$. In addition, it may well be that trends in true differentials are dominated by changes in $H_W(t) - H_{NW}(t)$ over time so that examination of observed wages obscures the true picture.

Empirical verification of the propositions outlined above may be obtained. Following the method described in my 1976 paper, a wage growth equation can be estimated so that one can ascertain $f(t)$ and $f'(t)$. Once that is done it is a simple matter to estimate

[5]More will be said on this in the empirical section.

[6]It is generally argued that age-earnings profiles are steeper for whites than nonwhites.

"true" wages by the approximation of $F(t)$ given in equation (4).

II. Estimation

Theory (see Gary Becker) and empirical evidence tell us that wage growth through *OJT* is larger for younger individuals than it is for older ones.[7] Age-earnings profiles are generally observed to be steeper in the early part of life so that $f'(t)$ is a decreasing function of t. Therefore differences between measured wages and true wages are likely to be greatest during the first years of work experience. For this reason, the two data sets to be used provide information exclusively on young individuals. The 1966–69 period is analyzed with the use of the *National Longitudinal Survey for Young Men* (*NLS*) 14–24 years of age in 1966. The 1972–74 period makes use of the *National Longitudinal Study of the High School Class of 1972* (*NLSHS*). The first empirical part of this section will be devoted to examination of the 1966–69 *NLS*. In the second part I will compare results from this earlier period to those obtained for the later 1972–74 period.

In order to estimate $f'(t)$ and equation (4), it is necessary to estimate a wage growth equation. A general form of wage growth equation is

$$W_\tau = W_t e^{\gamma(\tau - t)} \tag{5}$$

The growth rate γ depends upon many factors, the most important of which for this analysis is work experience. Thus, rewrite (5) as

$$\begin{aligned} W_\tau = W_t \exp [&\beta_0 + \beta_1 E_t + \beta_2 A_t \\ &+ \beta_3 \Delta H + \beta_4 \Delta S + \beta_5 \Delta ST \\ &+ \beta_6 \Delta E + \beta_7 U_\tau + \beta_8 U_\tau \Delta E \\ &+ \beta_9 M_t + \beta_{10} M_\tau + \beta_{11} D \\ &+ \beta_{12} D(\Delta E) + \beta_{13}(\Delta S)\Delta E] \end{aligned} \tag{6}$$

[7]There are some exceptions to this. Alan Blinder and Yoram Weiss, for example, provide a theoretical framework in which "recycling" is possible so that *OJT* investment may well decline and then increase again. Empirical counterexamples exist as well. These are viewed as exceptions rather than the rule.

TABLE 1

Variable	Definition
E_t	years of work experience by period t
W_τ	the hourly wage rate in cents in period τ
W_t	the hourly wage rate in cents in period t
A_t	the individual's age in period t
ΔH	the change in "usual" hours worked between t and τ
ΔS	the change in the highest grade of formal schooling completed between t and τ
ΔST	the difference between a dummy set equal to 1 if the individual was attending school in t and a dummy similarly defined[a] for τ
ΔE	the number of weeks worked between t and τ divided by fifty-two (i.e., it is the proportion of years worked)[b]
U_τ	a dummy set equal to 1 if the individual was in a union in τ
M_τ	a dummy set equal to 1 if the individual was married in τ
M_t	a dummy set equal to 1 if the individual was married in t
D	a dummy set equal to 1 if the individual is white
S_t	the highest grade of schooling completed by period t

[a]See my 1977 paper for the rationale behind this variable.

[b]This variable measures actual work experience time as distinguished from chronological time between period t and period τ.

Definitions are contained in Table 1.

A detailed discussion of the motivation behind this equation is provided elsewhere (see my 1976, 1977 papers) and will not be repeated here. Suffice it to say that this is a standard wage-generating function that can be found in very similar form throughout the literature on estimation of micro-wage functions.

Note that the t in equations (1)–(4) refers to experience time rather than chronological time. That is, the $f'(t)$ that is relevant is that wage growth which occurs as the result of job experience per se. It is only this component of growth that can be counted as part of the wage rate. Residual wage growth that occurs as one "ages," even in the absence of job experience, is not part of work compensation and as such should not be counted in the total return. Thus,

$$(7)\quad f'(t) = \frac{\partial W_\tau}{\partial \Delta E} = W_t[\exp(\beta_0 + \ldots + \beta_{13}(\Delta S)\Delta E)](\beta_6 + \beta_8 U_\tau + \beta_{12} D + \beta_{13}\Delta S) = W_\tau(\beta_6 + \beta_8 U_\tau + \beta_{12} D + \beta_{13}\Delta S)$$

Equation (7) is derivable from estimates of the coefficients in (6). These coefficients can be estimated by taking the *log* of both sides of (6) and bringing all wage variables to the left-hand side. Or,

$$(8)\quad \ln W_\tau - \ln W_t = \beta_0 + \beta_1 E_t + \beta_2 A_t + \beta_3 \Delta H + \beta_4 \Delta S + \beta_5 \Delta ST + \beta_6 \Delta E + \beta_7 U_\tau + \beta_8 U_\tau \Delta E + \beta_9 M_t + \beta_{10} M_\tau + \beta_{11} D + \beta_{12} D(\Delta E) + \beta_{13}(\Delta S)(\Delta E)$$

Note that this specification has the advantage that it eliminates unobserved ability components which affect wages similarly in both years. (That is, it is equivalent to allowing each individual to have his own constant term in a *log* wage levels equation.) This implies that the estimates obtained from (8) are not as likely to be biased by omitted ability variables.

A. *The National Longitudinal Survey, 1966–69*

This data set provides detailed information on schooling and work experience. The fact that it is longitudinal allows one to separate experience from vintage effects and further to determine more precisely the nature and extent of work experience during a particular period. In this section, t is 1966 and τ is 1969. Observations were dropped for which no wage rate in either year was reported or for which information was incomplete. The results of estimating equation (8) for this period are contained below:

$$(9)\quad \ln W_{69} - \ln W_{66} = \underset{(.125)}{.586} - \underset{(.00516)}{.01445E_{66}} - \underset{(.00480)}{.02097A_{66}} + \underset{(.00067)}{.00178\Delta H} + \underset{(.03284)}{.06288\Delta S} - \underset{(.02176)}{.14681\Delta ST} + \underset{(.03707)}{.09286\Delta E} + \underset{(.1016)}{.50531U_{69}} - \underset{(.03844)}{.13321U_{69}\Delta E} - \underset{(.02818)}{.09688M_{66}} + \underset{(.02449)}{.01287M_{69}} - \underset{(.0753)}{.08262D} + \underset{(.03049)}{.01294D(\Delta E)} - \underset{(.01341)}{.00615(\Delta S)\Delta E}$$

$N = 2115$; $R^2 = .18$; $S.E.E. = .4332$

(Standard errors are enclosed in parentheses.)

The interpretation of these coefficients is provided elsewhere (see my 1976 paper). The important number for the purpose at hand is $\partial W_\tau/\partial \Delta E$. Substituting the obtained coefficients into equation (7), one has

$$(10)\quad \left.\frac{\partial W_{66}}{\partial \Delta E}\right|_{D=1} = \overline{W}_{66}|_{D=1}(.09286 - .13321\,\overline{U}|_{D=1} + .01294 - .00615\,\overline{\Delta S}|_{D=1}) = (\$2.14)\underset{(.02280)}{(.06717)} = \underset{(.049)}{\$0.144}$$

for the mean white in the sample and

$$(11)\quad \left.\frac{\partial W_{66}}{\partial \Delta E}\right|_{D=0} = \overline{W}_{66}|_{D=0}(.09286 - .13321\,\overline{U}|_{D=0} - .00615\,\overline{\Delta S}|_{D=0}) = (\$1.57)\underset{(.02985)}{(.05589)} = \underset{(.04586)}{\$0.088}$$

for the mean black. (Standard errors are enclosed in parentheses below the coefficient).

From this, mean "actual" wages can be calculated by substituting into (4). Thus,

$$(12a)\quad \overline{F}(1966)|_{D=1} = \bar{f}(1966)|_{D=1} + .144\int_0^{45} e^{-.1\delta}\,d\delta = \$2.14 + \underset{(.48)}{\$1.42} = \underset{(.48)}{\$3.56}$$

$$(12b)\quad \overline{F}(1966)|_{D=0} = \bar{f}(1966)|_{D=0} + .088\int_0^{45} e^{-.1\delta}\,d\delta = \$1.57 + \underset{(.46)}{\$.87} = \underset{(.46)}{\$2.44}$$

if retirement occurs in 2011 and $r = .1$. Note

that r is the real rate of interest and includes depreciation. (The hours worked per year term enters both sides of the equation and thus cancels.) So for young white males, 40 percent of compensation comes in the form of on-the-job training. That proportion is about 36 percent for nonwhites. What is important to note is that although the observed differential, $\bar{f}(1966)|_{D-1} - \bar{f}(1966)|_{D-0}$ is only 57¢, the true differential, $\bar{F}(1966)|_{D-1} - \bar{F}(1966)|_{D-0} = \$1.12$ (with a standard error equal to \$0.66). Even in ratios, $\bar{f}(1966)|_{D-1}/\bar{f}(1966)|_{D-0} = 1.37$ whereas $\bar{F}(1966)|_{D-1}/\bar{F}(1966)|_{D-0} = 1.46$. Thus, the observed wage differential understates the true wage differential because whites and nonwhites differ in the amount of *OJT* they receive.

We are quick to add, however, that the difference in estimated *OJT* across the two groups is not large in 1966. In fact, $\bar{H}(1966)|_{D-1} - \bar{H}(1966)|_{D-0} = \$0.55$ (with a standard error of \$0.67) so that the difference is relatively small and the standard error is quite large. The conclusion is then that for this early period, the measured wage differential understates the true differential, but the extent of that understatement is questionable. The *OJT* component is substantial for both groups, but the difference in the size of those components does not vary greatly across groups.

Two additional points should be made. First, the reader should note that the coefficient on $(U_\tau)(\Delta E)$ is negative, while that on U_τ is positive. Thus, it appears that young union workers receive substantially higher wages than nonunion workers with similar characteristics, but that their profiles are flatter. Experience increases wages less in the union sector.

Second, although $\overline{\Delta S}$ and $\bar{U}_\tau$ do not differ greatly across groups, $\bar{W}_{66}$ does. Thus, most of the differential in *OJT* between whites and nonwhites is attributable to the fact that whites have higher initial wages. But for the purposes of comparison here, one does not want to hold initial observed wage rates constant. To do so would eliminate the observed differential as well as the unobserved component. Our task is not to determine whether wage differentials reflect productivity differences as opposed to discrimination. Rather, we simply want to determine what those differences actually are when measured correctly and how they change over time.

Even the above calculation is likely to understate the true differential for two reasons, both of which have to do with differences in labor force behavior between whites and nonwhites. First, since whites are less likely to suffer unemployment than are nonwhites and since they typically retire at a later date (see William Bowen and T. Aldrich Finegan), the time period over which (4) is integrated should be longer for whites. Thus, the assumption that $T - t = 45$ for all individuals is likely to cause the estimate of the value of wage growth for whites relative to nonwhites to be understated. This effect is admittedly small.

Second, since in any given year a white worker is more likely to be employed than a nonwhite worker (once he begins full-time participation in the labor force), $F(t)|_{D-1}$ and $F(t)|_{D-0}$ might reasonably be weighted by the probability of obtaining that wage. This, however, to be conceptually correct, requires a great deal of information on the value of leisure. If the world were in equilibrium, the marginal value of a minute spent in leisure must equal the marginal value of a minute of employment; that is, on the margin the value of leisure equals the wage rate. However, as changes in time worked come in discrete blocks, the value of leisure of that discrete block of time must necessarily be smaller than earnings. (This is simply the result of the diminishing marginal rate of substitution between leisure and goods as leisure increases.) The amount by which it is smaller depends upon the utility function itself. Without information on the parameters of that function it is impossible to choose correct weights for white vs. nonwhite wages. This notwithstanding, the estimates of wage differentials obtained above are lower bounds to the true differentials since both of these labor force participation effects work in the direction of increasing the true differential.

A final point on white-nonwhite differences

in wage growth are in order. Despite all that has been said above, it is not the case that during 1966–69 young men's wages grew at a more rapid rate for whites than nonwhites. On the contrary,

$$\frac{W_{69} - W_{66}}{W_{66}}\Big|_{D=0} = .55$$

whereas $\frac{W_{69} - W_{66}}{W_{66}}\Big|_{D=1} = .47$

This difference is reflected in the coefficient on D which is negative, although insignificant.[8] This is quite consistent with the scenario outlined above. If firms have responded to government pressure to reduce white-nonwhite wage differentials by reducing the wage growth component for nonwhites relative to whites, initial observed wages will rise relatively for nonwhites. That is, since the group in question consists of young men who are predominantly at entry level jobs, pecuniary wages of nonwhites will rise on these jobs at a more rapid rate for nonwhites than whites. This is reflected in the negative coefficient on D. Yet once those initial jobs are obtained, nonwhites' experience will have a different effect on wage growth than whites' experience. In this early period, the differential effect appears to be the result of differences in initial conditions. (Alternatively, nonwhites may benefit more from aging per se.) In the later period to be analyzed below, the wage growth induced by differential experience occurs even in the absence of differing initial conditions.[9]

[8]Regressions estimated separately for whites and nonwhites did not yield a significantly different sum of squared residuals from the combined regression ($F(10, 2091) = 1.14$).

[9]The validity of assumptions made initially were tested empirically. Above, it was shown in (3) that $C'(t) \geq (f(t) - f(t^*))/(t^* - t)$. Using Mincer's (1974) figures (see his chart 4.4), one finds that weekly earnings for white males with twelve years of schooling rise from about \$70 at two years of experience to a peak of about \$140 at forty years of experience or in per hour terms

$$C'(t) \geq \frac{\$1.75 - \$3.50}{38} \geq -\$.046$$

if the work week is assumed to be forty hours. Since $f'(t)$ was estimated to be \$.144 for this group, by (3) $\phi'(t)$ could be no lower than \$.098 by these estimates. Furthermore, if it is assumed that nonwhite profiles rise half as quickly as whites, the bias for nonwhites could be no larger in absolute value than \$.023. This would imply a $\phi'(t)$ for whites of no less than \$.065. Thus, the point made above and relative magnitudes still hold. Even if $f'(t)$ is a biased estimate of $\phi'(t)$, that bias appears to be relatively small. More importantly, since the bias affects both whites and nonwhites, the bias in $f'|_{D=1} - f'|_{D=0}$ tends to be even smaller.

Specification changes were considered. First, the ΔE term was interacted with the total number of jobs held during the period. It has been suggested that experience acquired on one job for a given period has a different effect on wage growth than experience acquired while on a number of different jobs for that same length. The coefficient on this variable was $-.0102$ (*s.e.* 0.0921) so that its effect appears to be small and insignificant. In addition, an $(Age)(\Delta E)$ term was added to the regression. It also failed to enter significantly. Given that some of these individuals are currently enrolled in school, it is interesting to perform the analysis on nonstudents. Equation (8) was therefore rerun (deleting the ΔS, $\Delta S(\Delta E)$, and ΔST terms) on the nonstudent subsample. This reduced the number of observations to 1,021 and standard errors increased as expected. However, the primary findings remained the same. $\bar{F}(1966)|_{D=1} = \$4.46$ while $\bar{F}(1966)|_{D=0} = \$1.94$, yielding a true differential of \$2.52. The observed wage rates were $\bar{f}(1966)|_{D=1} = \$2.65$ and $\bar{f}(1966)|_{D=0} = \$1.74$ so that the observed differential was 91¢ per hour.

B. *The National Longitudinal Study of the High School Class of 1972*

In this section, longitudinal data from the *NLSHS* will be used to estimate the unobserved *OJT* component of earnings during the later period. The *NLSHS* is a national probability sample of about 22,000 high school seniors. A survey was taken during the Spring of 1972 and two follow-ups were conducted in October of 1973 and 1974. For the purposes of this analysis, a subsample of males who had wage rates reported and who supplied complete information on the other relevant variables was selected (2,397 individuals fit this category). This sample, although similar to the *NLS* for 1966–69, has some important differences. The major difference is that all individuals in this sample were enrolled in the twelfth grade in 1972 so no early high school dropouts are contained. This also implies that the age distribution of respondents is much more tightly centered around the mean age in

1972 than was the 1966–69 sample (although the difference in mean ages is not that large). Furthermore, there is virtually no variation in initial schooling levels during October 1972, the date of the initial wage rate.

For this section, equation (8) is estimated where t refers to 1972 and τ to 1974. The results of that estimation are

$$(13)\quad \ln W_{74} - \ln W_{72} = \underset{(.278)}{.1248} + \underset{(.01145)}{.04174E_{72}} - \underset{(.01419)}{.000323A_{72}} - \underset{(.000712)}{.004232\Delta H} + \underset{(.08594)}{.23592\Delta S} - \underset{(.02345)}{.07094\Delta ST} + \underset{(.05357)}{.12840\Delta E} + \underset{(.0251)}{.02750U_{74}} + \underset{(.1410)}{.02522U_{74}(\Delta E)} - \underset{(.04899)}{.04593M_{72}} + \underset{(.02152)}{.04653M_{74}} - \underset{(.1111)}{.15013D} + \underset{(.06198)}{.08477D(\Delta E)} - \underset{(.04762)}{.16140(\Delta S)\Delta E}$$

Again, upon substituting the obtained coefficients into equation (7), one has

$$(14)\quad \frac{\partial W_{74}}{\partial \Delta E}\bigg|_{D=1} = (\overline{W}_{74}|_{D=1}(.1284 + .02522\overline{U}_{74}|_{D=1} + .08477 - .16140\,\overline{\Delta S}|_{D=1})) = \underset{(.03658)}{\$3.70(.16440)} = \underset{(.135)}{\$0.609}$$

for the mean white in the sample and

$$(15)\quad \frac{\partial W_{74}}{\partial \Delta E}\bigg|_{D=0} = (\overline{W}_{74}|_{D=0}(.1284 + .02522\overline{U}_{74}|_{D=0} - .16140\,\overline{\Delta S}|_{D=0})) = \underset{(.05104)}{\$3.69(.09093)} = \underset{(.188)}{\$0.335}$$

for the mean black. We then calculate mean actual wages as

$$(16)\quad \overline{F}(1974)|_{D=1} = \bar{f}(1974)|_{D=1} + .609\int_0^{45} e^{-.10}\,d\delta = \$3.70 + \underset{(1.33)}{\$6.02} = \underset{(1.33)}{\$9.72}$$

$$(17)\quad \overline{F}(1974)|_{D=0} = \bar{f}(1974)|_{D=0} + .335\int_0^{45} e^{-.10}\,d\delta = \$3.69 + \underset{(1.86)}{\$3.32} = \underset{(1.86)}{\$7.01}$$

Note that the difference between the *OJT* component for whites and nonwhites is now \$2.70 (*s.e.* 2.28) as opposed to \$0.55 (*s.e.* 0.67) for the earlier period. Although the standard error is relatively large, it appears as though there has been a large increase in the amount of *OJT* that whites receive relative to blacks. This shows up in the calculation of the actual wage differential. If we assume a 5 percent rate of inflation, then the \$2.71 differential for 1974 becomes \$1.81 in 1966 dollars. Thus, the change in the observed wage differential corrected for inflation is

$$[\bar{f}(1974)|_{D=1} - \bar{f}(1974)|_{D=0}]e^{-.4} - [\bar{f}(1966)|_{D=1} - \bar{f}(1966)|_{D=0}] = (.01)(.67) - (.57) = -\$.56$$

The observed differential fell to almost zero by 1974 so the difference in observed differentials is −56¢. The change in the true differential, however, is

$$[\overline{F}(1974)|_{D=1} - \overline{F}(1974)|_{D=0}]e^{-.4} - [\overline{F}(1966)|_{D=1} - \overline{F}(1966)|_{D=0}] = \$1.81 - \$1.12 = \underset{(2.39)}{\$0.69}$$

Thus, although the observed white-nonwhite differential seems to have fallen to zero over the eight-year period, the true differential, which includes the value of human capital, appears to have risen, if anything.[10]

Even holding other things constant, the

[10]This estimate is imprecise, so functional forms of wage growth equations other than the semilog specified in equation (8) were tried. First, the *log*-difference dependent variable was replaced with a linear difference. The results were that

$$[\overline{F}(1974)|_{D=1} - \overline{F}(1975)|_{D=0}]e^{-.4} - [\overline{F}(1966)|_{D=1} - \overline{F}(1966)|_{D=0}] = \$1.87 - 1.32 = \$0.55$$

This is virtually the same as the 69¢ estimated above.

coefficient on $D(\Delta E)$ is positive and substantial (although the standard error is still relatively large) for 1974. For 1966, this coefficient was essentially zero. The difference between the coefficients over the two periods is .07184 (*s.e.* 0.06907). This suggests a widening in the *OJT* gap which was concurrent with the narrowing of the pecuniary wage gap.[11]

The most important conclusion of this section then is that nonwhite gains in pecuniary wages over the eight-year period under study were more than offset by declines in the unobserved *OJT* component of earnings. It is also the case that in terms of levels of *OJT*, whites seem to receive substantially more than nonwhites in both periods. It is the change over time, however, that finds whites enjoying even greater gains in *OJT* than nonwhites. This causes the true differential to rise while the observed one falls.

C. *Some Useful Comparisons*

Before reaching the firm conclusion that the narrowing of the pecuniary differential has been offset by increases in the *OJT* differential, some points should be made.

First, it is important to note that 1966–69 was a period of rapid economic activity while 1972–74 was a recession. (The unemployment rate for white males 16 years of age and older averaged 2.6 percent between 1966–69 and 4.2 percent between 1972–74.) This has major implications for the estimates obtained.

Note, for example, that the coefficient on initial experience E_{66} and E_{72} bounces from negative to positive. The negative coefficient on E_{66} is explained elsewhere (1976) as reflecting the fact that *OJT* is acquired during the first years at work. The positive coefficient on E_{72} is likely to capture the fact that given initial wage rates, more senior workers are less affected by recessions than are their more junior counterparts. If this is the explanation, the effect is sufficient to offset the tendency to invest more during early years. Similarly, ΔH had a positive sign in the 1966–69 regression, but a negative one later. The reversal may be due to simultaneity bias brought on by recessionary changes. If the recession lowers the wages of some workers, in the short run they may increase their hours worked. This finding has shown up in past work (see Mincer, 1962, for example). Thus, the negative coefficient of ΔH might reflect the fact that those individuals who experience relative declines in wages also increase their hours worked during recessions.

Another difference across periods which may result from differences in business activity relates to the levels of the *OJT* component. The absolute magnitude of this compensation is much larger between 1972 and 1974 than it was in the earlier period. There are at least two possible explanations. First, along the lines of Richard Butler and Heckman, the differential attractiveness of welfare payments to low- vs. high-income individuals will tend to cause only the most able workers to remain. This is the case especially during recessions (since more drop below the point at which this alternative source of income is acceptable) so that the 1972–74 sample of individuals who are working is of higher average ability than the sample of workers during 1966–69. (They are also more able in that they all have attended school through grade twelve.) If, as is likely to be the case, the more able invest more in *OJT*, this would show up as a higher *OJT* component for the mean working individual between 1972 and 1974. Second, I have argued elsewhere (1974) that it is rational to acquire *OJT* to a greater extent during recessions than during expansions. If so, the difference between the two periods would be a manifestation of this phenomenon. Although an interesting theoretical insight, Freeman (1978) cautions that there are serious data problems in the Butler and Heckman analysis so that it is perhaps an unjustifiably easy reconciliation for the puzzle.

Other findings may be reconciled. Welch has argued that black schooling quality has gone up over the 1960's and that this has resulted in an increase in the black-white

[11]Pooled regressions were run where each year's change in wages was analyzed as a separation observation. The results, which were consistent with the grouped data, are obtainable from me upon request.

earnings ratio. Freeman (1975) suggests that the growth was the result of the increased role of blacks in the political process which resulted in institutional changes causing blacks' wages to rise. The studies use wage rates or annual earnings of the individuals to examine differentials. But particularly for young workers, ignoring the human capital or wage growth component of earnings may seriously disguise wage differentials and their trends. Indeed, this may help explain why Freeman finds a narrowing of the differential for younger workers that is absent among older cohorts. Since older workers have flatter profiles, there is less room to take back in wage growth what has been given in pecuniary wage levels. Therefore, employer resistance to elimination of pecuniary differentials is likely to be greater with respect to older workers. Further, the fact that pecuniary wage differentials have narrowed to a greater extent for the highly educated as found by Freeman can be explained by these findings. If education and *OJT* are complementary, highly educated nonwhites would have a larger *OJT* component in 1966 than would the less educated. If, say, the same proportion of *OJT* was reduced for all nonwhites, pecuniary wages would rise by more in absolute terms for the highly educated. This would imply a greater narrowing of the differential for this group. Note that the narrowing of wage differentials to a larger extent for educated groups is consistent with Butler and Heckman. The explanation reconciles their story with the observed result since true differentials do not seem to have narrowed at all.

A final point is that the R^2 on the 1972–74 regression is considerably lower than the R^2 for 1966–69. The later data set deals with a much more homogeneous group than the former. (There is no variation in initial schooling and very little in age.) Thus, regressions on the *NLSHS* are asked to perform a much more difficult task than are those on the original *NLS*.[12]

The most important conclusion is that although black-white pecuniary wage differentials were eliminated between 1966 and 1974, there is evidence to suggest that this tremendous improvement in the relative position of blacks may be, at least in part, illusory. If so, newer cohorts of blacks will have higher initial wage levels, but flatter profiles relative to whites than did previous cohorts of similar workers.

---

[12]Some checks on functional form were also performed for the 1972–74 period. First, the 1972–74 regression was performed for the nonstudent subsample. Again ΔS, ΔST, and $(\Delta S)(\Delta E)$ were deleted. Standard errors increased, but the coefficient on $D(\Delta E)$ was still positive and substantial at .09002 (*s e* 0.07988). Furthermore, $\bar{F}(1974)|_{D=1}$ = \$10.06 and $\bar{F}(1974)|_{D=0}$ = \$6.46 yielding a true differential of \$3.60 per hour. The observed differential was only \$2.64 – \$2.56 or 8¢ per hour. It should also be noted that the \$2.52 differential obtained for the nonstudent group in 1966, when inflated at 5 percent per year, becomes \$3.75 in 1974 dollars. This compares quite favorably to the estimated \$3.60 differential for 1974. The test for an $(Age)(\Delta E)$ interaction was performed. Neither age, age (ΔE), nor the two taken jointly entered significantly. Nor did stratification of the sample into two separate groups on the basis of age yield significantly different results. Evidence on the difference between observed and true differentials for males and females is provided in my 1978 paper.

REFERENCES

O. Ashenfelter and G. Johnson, "Unionism, Relative Wages and Labor Quality in U.S. Manufacturing Industry," *Int. Econ. Rev.*, Oct. 1972, *13*, 488–508.

Gary S. Becker, *Human Capital*, New York 1964; 2d ed. 1975.

Y. Ben-Porath, "The Production of Human Capital and the Life-Cycle Model of Labor Supply," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

A. Blinder and Y. Weiss, "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ.*, June 1976, *84*, 449–72.

William G. Bowen and T. Aldrich Finegan, *The Economics of Labor Force Participation*, Princeton 1970.

R. Butler and J. J. Heckman, "The Impact of the Government on the Labor Market Status of Black Americans: A Critical Review of the Literature and Some New Evidence," unpublished paper, Univ. Chicago 1976.

Richard Freeman, *The Over-Educated American*, New York 1976.

———, "Socio-Economic Mobility and Black-White Economic Differences in the

'New Market' for Black Labor," unpublished paper, Harvard Univ. 1976

———, "Time Series Evidence on Black Economic Progress: Shifts in Demand or in Supply?," disc. paper no. 6321, Harvard Inst. Econ. Res. 1978.

———, *The U.S. Discriminatory System*, forthcoming.

W. Haley, "Human Capital: The Choice Between Investment and Income," *Amer. Econ. Rev.*, Dec. 1973, *63*, 929–44.

J. J. Heckman, "Estimates of Human Capital Production Function Embedded in a Life Cycle Model of Labor Supply," in Nestor Terleckyj, ed., *Household Production and Consumption*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, New York 1975.

P. G. Keat, "Long-Run Changes in Occupational Wage Structure, 1900–1956," *J. Polit. Econ.*, Dec. 1960, *67*, 584–600.

E. P. Lazear, "The Timing of Technical Change: An Analysis of Cyclical Variations in Technology Production," unpublished doctoral dissertation, Harvard Univ. 1974.

———, "Human Wealth and Human Capital," Nat. Bur. Econ. Res. work. paper no. 97, Stanford 1975.

———, "Age, Experience and Wage Growth," *Amer. Econ. Rev.*, Sept. 1976, *66*, 548–58.

———, "Schooling as a Wage Depressant," *J. Human Resources*, Spring 1977, *12*, 164–76.

———, "Male-Female Wage Differentials: Has the Government Had Any Effect?," in Cynthia Lloyd, ed., *Women in the Labor Market*, New York 1978.

H. Gregg Lewis, *Unionism in Relative Wage in the United States*, Chicago 1963.

L. Lillard, "Inequality: Earnings vs. Human Wealth," *Amer. Econ. Rev.*, Mar. 1977, *67*, 42–53.

Jacob Mincer, "Labor Supply, Family Income, and Consumption," *Amer. Econ. Rev. Proc.*, May 1960, *50*, 574–83.

———, "Labor Force Participation of Married Women," in *Aspects of Labor Economics*, Universities–Nat. Bur. Econ. Res. conference series, Princeton 1962.

———, *Schooling, Experience and Earnings*, New York 1974.

M. W. Reder, "Theory of Occupational Wage Differentials," *Amer. Econ. Rev.*, Dec. 1955, *45*, 833–52.

———, "Wage Differentials: Theory and Measurement," in *Aspects of Labor Economics*, Universities–Nat. Bur. Econ. Res. conference series, Princeton 1962.

S. Rosen, "Trade Union Power, Threat Effects, and the Extent of Organization," *Rev. Econ. Stud.*, Apr. 1969, *36*, 185–96.

———, "Unionism and Occupational Structure," *Int. Econ. Rev.*, June 1970, *11*, 269–86.

———, "Income Generating Functions and Capital Accumulation," Nat. Bur. Econ. Res. disc. paper no. 306, Harvard Univ. 1973.

J. P. Smith and F. Welch, "Race Differences in Earnings: A Survey and New Evidence," R-2295-NSF, Rand Corp., Mar. 1978.

F. Welch, "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *63*, 893–907.

National Longitudinal Survey for Young Men, 1966–1969, (*NLS*) Ohio State Univ.

National Longitudinal Study of the High School Class of 1972, 1972–1974, (*NLSHS*) Research Triangle Institute, Chapel Hill.

Models of the Firm and International Trade under Uncertainty

By DAVID P. BARON AND ROBERT FORSYTHE*

One of the significant advances in economic theory has been the incorporation of uncertainty into the models used to investigate economic behavior. The explicit treatment of uncertainty has permitted economists to predict the behavior of economic agents operating in an uncertain environment and to explain, for example, the existence of insurance, stock markets, and forward exchange markets that have no necessary role in a deterministic world. One natural application of the economics of uncertainty has been to the study of international trade and exchange in which uncertainty regarding exchange rates and relative prices is a prominent feature of the environment of economic agents. The purpose of this paper is to frame the international trade results developed in the recent works of Wolfgang Mayer and Raveendra Batra in light of the current state of the theory of the firm under uncertainty. Before analyzing the effect of uncertainty on international trade, a perspective on the application of the economics of uncertainty to neoclassical theory will be presented with an emphasis on the theory of the firm.

One class of models into which uncertainty has been incorporated can be labeled as "entrepreneurial" models in which the firm is assumed to maximize the expected utility of profit.[1] The results from these models indicate, for example, that production decisions depend on the preferences and expectations of the entrepreneur and thus that the results of deterministic theory do not in general obtain. The question that remains, however, is whether these models are appropriate only for a firm owned by an entrepreneur or if they pertain to firms that are owned by more than one individual.

A second class of models assumes that markets exist such that contracts either can be traded contingent on each state of nature or such that there are as many securities as there are states.[2] The principal result of these "complete market" models is that consumers have sufficient opportunities to trade so that marginal rates of substitution are equated across states. In this case, the owners of the firm are unanimous in preferring that the firm maximize its market value using the established market prices to evaluate alternative production plans, and hence, the results of deterministic models obtain. When markets, however, are incomplete in the sense that there are more states than there are securities that can be traded, value maximization may not be in the best interests of the owners of a firm. Furthermore, the owners may not agree on an objective for a firm, since in an incomplete market consumers' marginal rates of substitution cannot be equated through their opportunities to trade.

A third class of models that deals with these issues was initiated by Peter Diamond's study of a model with an incomplete market structure in which the only opportunity consumers have to allocate risks is by trading the shares of firms in a securities market. Even though consumers are unable to equate their marginal rates of substitution in each state, they do equate their marginal rates of substitution between every pair of securities. Then, assume that the vectors (across states) of marginal returns from a production plan of a firm are contained in the subspace spanned by the return vectors of firms. Satisfaction of this spanning condition permits use of the marginal rates of substitution between securities to demonstrate that shareholders unanimously prefer that the firm maximize its market value. Applying this theory to a model

*Graduate School of Management, Northwestern University, and California Institute of Technology, respectively.

[1]See Baron for an early survey of this class of model and Hayne Leland for a further contribution.

[2]See Kenneth Arrow and Gerard Debreu for the principal results for this class of model.

in which the firm is privately held but its securities are publicly traded, the usual results of deterministic theory are obtained with a securities market established certainty-equivalent used in place of the market prices that would be present in a complete market.

A complication in the use of these market models involves the separation of ownership from control. This separation raises the possibility that managers may believe that they are better informed than shareholders and may use their own judgements in evaluating alternative production plans. The incomplete market model to be considered here will be analyzed first under the assumption that the managers of a firm act in the best interests of the firm's shareholders, and then under the assumption that managers use their own or some chosen preferences and expectations in directing the firm.

Batra and Sandwip Das have considered an entrepreneurial model of a firm engaged in international trade, in which the firm maximizes the expected utility of profit in the presence of technological uncertainty. They conclude that risk aversion invalidates both the Heckscher-Ohlin and Rybczynski theorems. Mayer has analyzed an entrepreneurial model in which final commodity prices are uncertain. He demonstrated that with representative identical firms and free and costless entry, the Rybczynski and the Stolper-Samuelson theorems hold if the theorems are restated in terms of a "change in expected price, with higher central moments constant" (p. 797). He also concludes that the Factor-Price Equalization theorem obtains if the "utility functions and probability assessments of a given industry's firms are identical in both countries" (p. 804). These results are not contradictory in the context of entrepreneurial models, since Batra's model pertains to the "intermediate run," in Mayer's terminology, while Mayer considers an industry in long-run equilibrium.[3]

Elhanan Helpman and Assaf Razin (1976a,b; 1978) have shown in the context of an incomplete market model that the opportunity to trade the shares of firms in a securities market is sufficient to yield all of the standard results of international trade. The model to be considered here will be analyzed both in the context of an incomplete securities market and of an extension of the entrepreneurial class of models. In the incomplete market model the firm will be assumed to act in the best interest of its shareholders, and shareholders will be shown to unanimously prefer that the firm acts to maximize its market value. Batra and Mayer assume, however, that the firm maximizes the expected utility of profit for some utility function and expectations, and find that in the intermediate run the behavior of the firm depends on the chosen utility function and expectations. The analogous assumption within the context of the model considered here is that because of the separation of ownership from control, the manager of the firm uses some chosen preferences and expectations to guide the firm. The second assumption under which the model will be analyzed is thus that the "manager cum entrepreneur" maximizes the expected utility of the total residual profits that accrue to shareholders after the purchase of treasury shares for an arbitrary strictly concave utility function and arbitrary expectations. The result in this case is that the firm will be operated in exactly the same manner as if it were operated directly in the best interests of shareholders. Furthermore, under either assumption the market equilibrium has the same properties as in a deterministic model, and all the standard theorems of international trade hold. Consequently, the reformulated entrepreneurial model leads to the same behavior as does the incomplete market model.

I. The Model

To examine the Batra and Mayer models, consider a two-sector model in which a firm in sector 1 is characterized by a technology utilizing factor inputs K_1 and L_1 of capital and labor, respectively, to produce an uncertain output X_1 of commodity 1 given by

$$X_1 = \alpha F_1(K_1, L_1)$$

where F_1 is a production function and α is a

[3]See Murray Kemp for concise statements of the theorems of international trade.

random variable, $\alpha \geq 0$. The output X_2 of commodity 2 produced by a firm in the second sector is deterministic and given by

$$X_2 = F_2(K_2, L_2)$$

Each sector is assumed to be composed of many firms, although only one firm in each sector will be designated and analyzed. The production functions F_1 and F_2 are assumed to be linear homogeneous and strictly concave so that

$$F_i(K_i, L_i) = L_i f_i(k_i), \qquad i = 1, 2$$

where k_i is the capital-labor ratio, $f_i' > 0$ and $f_i'' < 0$. The firms make their input decisions prior to observing α and are assumed to sell their output at the price determined in a perfectly competitive market. Letting the price of the output of the first sector expressed in terms of the second commodity be denoted by p,[4] the profit of a firm in sector 1 is

$$\pi_1 = p\alpha\, L_1 f_1(k_1) - wL_1 - rK_1$$

where w and r are the factor prices which are assumed to be determined in a competitive market.

Each firm's input decisions are made at the beginning of the period, α is then observed, and at the end of the period firms distribute their profits to their owners who then purchase commodities under certainty. Letting $U^i(C_1^i, C_2^i)$ denote consumer i's ordinal utility function for consumption of the two commodities at the end of the period, his consumption problem is to choose (C_1^i, C_2^i) given his income $I^i(\alpha^0)$

$$\text{to maximize}_{C_1^i, C_2^i} \; U^i(C_1^i, C_2^i)$$

$$\text{subject to } pC_1^i + C_2^i \leq I^i(\alpha^0)$$

where α^0 is a realization of α. The ordinary demand functions will be denoted by $C_1^i(I^i(\alpha^0), p)$ and $C_2^i(I^i(\alpha^0), p)$. Since at the end of the period when consumption decisions are made the supplies of the two commodities are fixed at X_1 and X_2, the equilibrium price $p(\alpha^0)$ is determined by the solution to the market-clearing conditions

$$\sum_i C_1^i(I^i(\alpha^0), p(\alpha^0)) = X_1$$

$$\sum_i C_2^i(I^i(\alpha^0), p(\alpha^0)) = X_2$$

Consequently, if the output of the first sector is uncertain, the relative price must be uncertain as indicated by Helpman and Razin (1976a). Batra assumed that the output price is constant when supply is uncertain but such an assumption is unwarranted. Furthermore, as will be demonstrated below, an uncertain price does not affect the standard results of international trade. Mayer considered an uncertain price assuming that it is determined "internationally."

The solution to the consumer's end-of-period consumption problem determines his indirect utility function denoted by $U^i(I^i(\alpha), p(\alpha))$ which the consumer will use in making portfolio decisions. The uncertain price appears both as an argument of the utility function and as a determinant of income $I^i(\alpha)$, since the profit of firms depends on that price. The indirect utility function is assumed to be strictly concave in $I^i(\alpha)$, indicating risk aversion, and will be used to characterize the consumer's investment behavior.

A consumer i is assumed to be endowed with fixed quantities of labor L^i and capital K^i that can be sold to any firm at prices w and r, respectively. A consumer is also endowed with cash $\bar{y}^i$ and an ownership share in firm j denoted by $\bar{\gamma}_j^i$, where $\Sigma_i \bar{\gamma}_j^i = 1$, $j = 1, 2$. Consumers may sell their initial ownership shares or can purchase new shares γ_j^i in a securities market, which at the beginning of the period establishes the market values V_1 and V_2 of the two firms. The income $I^i(\alpha)$ available for consumption is obtained from savings y^i,[5] from the sale of labor and capital, and from share ownership which entitles the consumer to a share of profits or

$$I^i(\alpha) = \gamma_1^i \pi_1 + \gamma_2^i \pi_2 + y^i + wL^i + rK^i$$

[4] Batra uses p to denote the relative price of commodity 2 in terms of commodity 1. The alternative definition is used here because, as will be indicated, the price will be a random variable, and we wish to place all the uncertainty in the first sector.

[5] The interest rate is assumed to be zero. A positive interest rate will not alter the results.

Consumers are assumed to have expectations regarding $p\alpha$ that are expressed as a distribution function $G^i(p,\alpha)$, and expectations are assumed to be independent of the allocations made at the beginning of the period. That is, consumers and firms are small enough that their actions are not perceived to affect the relative commodity price.

At the beginning of the period the consumer has the portfolio problem

$$\text{to maximize}_{\gamma_1^i, \gamma_2^i, y^i} \; E^i U^i(I^i(\alpha), p(\alpha))$$

$$\text{subject to } y^i + \gamma_1^i V_1 + \gamma_2^i V_2 \leq \bar{y}^i + \bar{\gamma}_1^i V_1 + \bar{\gamma}_2^i V_2$$

where $\bar{\gamma}_1^i V_1 + \bar{\gamma}_2^i V_2$ is the value of the initial endowment of shares and $(\gamma_1^i V_1 + \gamma_2^i V_2)$ is the cost of purchasing the new portfolio. Solving the budget constraint for savings y^i and substituting into $I^i(\alpha)$, the portfolio optimality conditions are

$$E^i[U_1^i \cdot (\pi_1 - V_1)] = E^i[U_1^i \cdot (p\alpha L_1 f_1(k_1) - wL_1 - rL_1 k_1 - V_1)] = 0$$

$$E^i[U_1^i \cdot (\pi_2 - V_2)] = E^i[U_1^i] \cdot (L_2 f_2(k_2) - wL_2 - rL_2 k_2 - V_2) = 0$$

where U_1^i denotes $\partial U^i/\partial I^i(\alpha)$. Dividing the optimality conditions by $\int U_1^i \cdot dG^i(p\alpha)$ and defining the consumer's implicit price as

$$\rho^i(p,\alpha) = U_1^i \cdot g^i(p,\alpha)/(\int U_1^i \cdot dG^i(p,\alpha))$$

where $g^i(p,\alpha)$ is the density function corresponding to G^i, the optimality conditions can be rewritten as

$$(1) \quad (\int \rho^i(p,\alpha)(p\alpha)dpd\alpha) L_1 f_1(k_1) - wL_1 - rL_1 k_1 = V_1$$

$$(2) \quad L_2 f_2(k_2) - wL_2 - rL_2 k_2 = V_2$$

Since $\int \rho^i(p,\alpha)dpd\alpha = 1$, $\int V_1 \rho^i(p,\alpha)dpd\alpha = V_1$ and similarly for the other terms that do not depend on the realization of the random variables p and α. The implicit price $\rho^i(p^0,\alpha^0)$ is the amount consumer i would pay for a security that pays one dollar if and only if (p^0, α^0) obtains,[6] or equivalently, the marginal rate of substitution between a dollar if (p^0,α^0) obtains and a dollar obtained with certainty. A security such as cash that pays one dollar for any (p,α) outcome thus has a value of one dollar. A securities market equilibrium is assumed to exist.

The securities market ensures that the quantity $\int \rho^i(p,\alpha)p\alpha dpd\alpha$ will be the same for all consumers, since solving from (1) yields

$$(3) \quad \int \rho^i(p,\alpha)p\alpha dpd\alpha = (V_1 + wL_1 + rL_1 k_1)/(L_1 f_1(k_1))$$

The right-hand side is independent of i, so $\int \rho^i(p,\alpha)p\alpha dpd\alpha$ is the same for all i. This quantity may be interpreted as the market price for a unit of "certain" production $L_1 f_1(k_1)$, since the numerator is the payment to factor inputs plus the rent V_1 to owners and the denominator is the certainty portion of output. Since the term in (3) is the value of a unit of certain output $L_1 f_1(k_1)$, it will be called the market certainty-equivalent price and will be denoted by p^*.

Suppose that there are two countries each having firms in both sectors. In the presence of an international securities market the consumer's portfolio problem remains unchanged except for the fact that the ownership shares which may be purchased may now be distinguished by country. A representative firm in each sector may still be designated and, from (1) and (2), it can be seen that the opportunity for consumers to trade in shares in both countries enables the market mechanism to operate so as to equate the marginal rates of substitution for securities among consumers in both countries, and thus the market certainty-equivalent p^* is the same for firms in these countries.[7]

II. Firm Input Decisions

In order to determine the optimal inputs, it is necessary to specify an appropriate criterion for the firm. Batra and Mayer assume

[6]The securities market is incomplete in this model, because there are more states than there are securities. Consequently, the implicit price $\rho^i(p,\alpha)$ for consumer i may differ from the implicit price $\rho^j(p,\alpha)$ for consumer j. If the markets were complete, the implicit prices for every consumer would be equal, since each would be able to make trades contingent on every state (p,α).

[7]As is customary in models such as these, it is assumed that there is no exchange rate uncertainty.

that the firm acts so as to maximize the expected utility of profit for some chosen utility function and expectations. Such an assumption is clearly warranted if the preferences are those of an entrepreneur who is the sole owner of the firm, but most firms are owned by shareholders and not by a single entrepreneur. Investor-owned firms are in principle to be operated in the best interests of their shareholders, but the separation of ownership from control suggests that the managers of firms may base their decision on their own preferences or on preferences that the manager believes are appropriate for his shareholders. Both the case in which the firm is operated in the interests of shareholders and the case in which the manager chooses some specific preferences to represent his or the shareholders' preferences will be considered. Given competitive behavior, the input decisions in both cases will be shown to be the same and to be functions only of prices observable in the markets. This is in contrast to the results of Batra and Mayer in which the optimal inputs depend on the preferences and expectations used.

A. *The Shareholder Interests Criterion*

With this criterion firms are assumed to act in the best interests of their shareholders. To determine a shareholder's preferences for the level of a firm's inputs, consider variations in consumer i's expected utility evaluated at a securities market equilibrium. Differentiating the expected utility evaluated at the optimal portfolio $\hat{\gamma}^i_1$, $\hat{\gamma}^i_2$, and $\hat{y}^i$ yields

$$(4)\quad \frac{\partial E^i\overline{U}^i}{\partial k_1} = \hat{\gamma}^i_1\left[\left(\int \rho^i(p,\alpha)p\alpha dp d\alpha\right) L_1 f_1'(k_1) - rL_1\right] + (\overline{\gamma}^i_1 - \hat{\gamma}^i_1)\frac{\partial V^i_1}{\partial k_1}$$

$$(5)\quad \frac{\partial E^i\overline{U}^i}{\partial L_1} = \hat{\gamma}^i_1\left[\left(\int \rho^i(p,\alpha)p\alpha dp d\alpha\right) f_1(k_1) - rk_1 - w\right] + (\overline{\gamma}^i_1 - \hat{\gamma}^i_1)\frac{\partial V^i_1}{\partial L_1}$$

where $\partial V^i_1/\partial L_1$ and $\partial V^i_1/\partial k_1$ are i's forecast of the change in the market value of the firm and $\partial E^i\overline{U}^i/\partial k_1 = (\partial E^iU^i/\partial k_1)/(EU^i_1)$ and similarly for L_1. Since the consumer knows that a securities market equilibrium will be established for any level of inputs contemplated, the forecasts may be determined by differentiating (1) to obtain

$$(6)\quad \left(\int \frac{\partial \rho^i(p,\alpha)}{\partial k_1} p\alpha dp d\alpha\right) L_1 f_1(k_1) + \left(\int \rho^i(p,\alpha)p\alpha dp d\alpha\right) L_1 f_1'(k_1) - rL_1 = \frac{\partial V^i_1}{\partial k_1}$$

$$(7)\quad \left(\int \frac{\partial \rho^i(p,\alpha)}{\partial L_1} p\alpha dp d\alpha\right) L_1 f_1(k_1) + \left(\int \rho^i(p,\alpha)p\alpha dp d\alpha\right) f_1(k_1) - rk_1 - w = \frac{\partial V^i_1}{\partial L_1}$$

As is usual in models in which there is an incomplete set of markets for risk sharing, each consumer will in general perceive a change in his implicit prices as given in the first term of (6) and (7). If however, changes in the inputs of one firm have a negligible effect on the availability of inputs of other firms and all consumers perceive that the profit and market value of a firm is independent of the decisions of any other firm, the change in the certainty-equivalent price can be shown to be zero. Since there are many firms in each industry, consumer i's forecast of the change in the market value of another firm in the first industry will be zero, which implies that

$$\int \frac{\partial \rho^i(p,\alpha)}{\partial k_1} p\alpha dp d\alpha = \frac{\partial p^*}{\partial k_1} = 0$$

$$\int \frac{\partial \rho^i(p,\alpha)}{\partial L_1} p\alpha dp d\alpha = \frac{\partial p^*}{\partial L_1} = 0$$

Thus each consumer may perceive a change in his implicit prices but this variation is constrained, since he acts as a price taker with respect to the market certainty-equivalent price p^*. This is analogous to the usual pure competition assumption.

With this result (6) and (7) may be substituted into (4) and (5) to obtain

$$(8)\quad \frac{\partial E^iU^i}{\partial k_1} = \overline{\gamma}^i_1 \frac{\partial V^i_1}{\partial k_1}$$

$$= \overline{\gamma}^i_1\left(\left(\int \rho^i(p,\alpha)p\alpha dp d\alpha\right) L_1 f_1'(k_1) - rL_1\right)$$

$$= \overline{\gamma}^i_1(p^* L_1 f_1'(k_1) - rL_1)$$

$$(9)\quad \frac{\partial E^iU^i}{\partial L_1} = \bar{\gamma}_1^i \frac{\partial V_1^i}{\partial L_1}$$
$$= \bar{\gamma}_1^i((\int \rho^i(p,\alpha)p\alpha dp d\alpha) f_1(k_1) - w - rk_1)$$
$$= \bar{\gamma}_1^i(p^* f_1(k_1) - w - rk_1)$$

All consumers who are initial shareholders ($\bar{\gamma}_1^i > 0$) will thus be unanimous with respect to a change in k_1 and L_1, since they each use the market certainty equivalent p^* given in (3) and thus the right sides of (8) and (9) depend only on shareholder characteristics through their ownership share.[8]

The factor input levels unanimously preferred by all initial shareholders may be determined from (8) and (9) and satisfy

$$(10)\quad p^* f_1'(k_1) - r = 0$$

$$(11)\quad p^* f_1(k_1) - rk_1 - w = 0$$

For the firm in sector 2 the analogous conditions are

$$(12)\quad f_2'(k_2) - r = 0$$

$$(13)\quad f_2(k_2) - rk_2 - w = 0$$

An equilibrium in the factor markets requires that the factor rewards be the same for both firms,

$$(14)\quad p^* f_1'(k_1) = f_2'(k_2)$$

$$(15)\quad p^* f_1(k_1) - k_1 f_1'(k_1)) = f_2(k_2) - k_2 f_2'(k_2)$$

and that resources are fully employed:

$$(16)\quad K = \sum_i K^i = L_1 k_1 + L_2 k_2$$

$$(17)\quad L = \sum_i L^i = L_1 + L_2$$

An equilibrium is assumed to exist such that positive amounts of both commodities are produced.

The conditions in (10) and (11) are identical to those for a firm in a deterministic world and do not depend on the characteristics of any consumer. This results because the securities market allows all consumers to make trades until the marginal return p^* for a dollar of investment per unit of certain output is the same. Furthermore, the equilibrium market value of the firm is zero, since multiplying (11) by L_1 yields

$$(18)\quad 0 = p^* f_1(k_1) - w - k_1 r = p^* L_1 f_1(k_1) - wL_1 - rL_1 k_1 = V_1$$

This is the same result as in the deterministic case for a competitive firm with a linear homogeneous production function.

The above demonstration indicates that Batra's intermediate-run case and Mayer's long-run equilibrium are identical for investor-owned firms in the sense that a competitive firm with a linear homogeneous production function has a zero market value. The opportunity for consumers to allocate their risks in a security market provides firms with the needed information p^* to plan their inputs efficiently, and constant returns to scale ensure that no firm may earn an excess return and thus that the market value is zero.

B. *A Managerial Model*

Separation of ownership from control of a firm suggests that managers may not wish to or may not be able to determine how the firm can be operated in the best interests of its shareholders. The manager then may choose some representative utility function and expectations to use in decision making. Since one alternative open to any firm is to purchase its own shares and hold them as treasury shares, the manager will be assumed in this "manager cum entrepreneur" model to maximize the expected utility of the net (after payment for treasury shares) cash flow per share that accrues to the outstanding shares. Viewing the payment for treasury shares as a cost, this is equivalent to maximizing the expected utility of earnings per share. Letting γ_1^* denote the percentage of the shares purchased by the firm and held in its treasury, the cash flow per share is $\pi_1^* = (\pi_1 - \gamma_1^* V_1)/(1 - \gamma_1^*)$. If the manager employs a concave utility function $U^*(\pi_1^*)$ and expecta-

[8]The same result obtains if each firm faces a "firm-specific" technological risk, since the vector of returns for a production plan of a firm is a multiplicative factor of the return vector for any other production plan. When the profit of a firm is not linear in the random variables, a market certainty equivalent cannot be determined as in (3).

tions denoted by the distribution function $G^*(p,\alpha)$, the first-order conditions for γ_1^*, L_1, and k_1 are

$$(19)\ E^*[U^{*\prime}\cdot(-V_1(1-\gamma_1^*)+\pi_1-\gamma_1^*V_1)] = E^*[U^{*\prime}\cdot(\pi_1-V_1)]=0$$

$$(20)\ E^*[U^{*\prime}\cdot(p\alpha f_1(k_1)-w-rk_1)]=0$$

$$(21)\ E^*[U^{*\prime}\cdot(p\alpha L_1 f_1'(k_1)-rL_1)]=0$$

where the market value of the firm is assumed to be unaffected by the firm's share purchases.[9] The manager's (or firm's) implicit prices $\rho^*(p,\alpha)$ may be defined as

$$\rho^*(p,\alpha) = U^{*\prime}\cdot g^*(p,\alpha)/(\int U^{*\prime}\cdot dG^*(p,\alpha))$$

and the manager's certainty equivalent is seen to be the same as that given in (3). Dividing (19), (20), and (21) by $E^*U^{*\prime}$ and substituting from (19) gives the conditions in (10) and (11), so the managerial model yields the same results as the model in the previous section.

The difference between the results of this model and that of Batra is that here the firm may purchase treasury shares, and this requires the manager to utilize the same market certainty-equivalent used by all shareholders. Consequently, the manager chooses the same levels of inputs as those that result from acting directly in the best interests of shareholders. Trading in the securities market allows the manager to align his marginal rate of substitution with that of all consumers, and hence, the optimal inputs depend only on market observable prices. Batra does not permit the manager to trade in a securities market, and thus, the optimal levels of k_1 and L_1 depend on the manager's preferences and expectations.

III. The Theorems of International Trade

Batra and Das conclude that when firms maximize their expected utility of profit the Heckscher-Ohlin and the Rybczynski theorems do not obtain. Mayer finds that these theorems hold when he assumes that entry is free and that all firms are identical, but he concludes that the Factor-Price Equalization theorem does not hold because "factor returns are crucially dependent on the utility functions and probability assessments of firms" (p. 803). When the shares of firms may be traded in a securities market, the factor returns depend only on the market observable certainty-equivalent as indicated in (10)–(13). The purpose of this section is to briefly indicate that the standard theorems of international trade obtain in this case.

The first step in the development is to show that there is a one-to-one relationship between the labor-capital factor-price ratio $\omega = w/r$ and the commodity-price ratio. From (10), (11), (12), and (13) the factor-price ratio is

$$(22)\quad \omega = f_i(k_i)/f_i'(k_i) - k_i, \qquad i=1,2$$

Differentiation yields

$$(23)\quad \frac{d\omega}{dk_i} = \frac{-f_i''(k_i)f_i(k_i)}{(f_i'(k_i))^2} > 0$$

which is the desired result. The factor rewards condition in (14) can be differentiated to obtain

$$(24)\quad \frac{dp^*}{d\omega}f_1'(k_1) + p^*f_1''(k_1)\frac{dk_1}{d\omega} = f_2''(k_2)\frac{dk_2}{d\omega}$$

so

$$(25)\quad \frac{dp^*}{d\omega}/p^* = \frac{f_2''}{f_2'}\frac{dk_2}{d\omega} - \frac{f_1''}{f_1'}\frac{dk_1}{d\omega} \quad \text{(substituting from (14))}$$

$$= -\frac{f_2'}{f_2} + \frac{f_1'}{f_1} \quad \text{(substituting from (23))}$$

$$= -\frac{1}{\omega+k_2} + \frac{1}{\omega+k_1} \quad \text{(substituting from (22))}$$

With the usual factor-intensity assumption it is evident that an increase in the factor-price ratio results in an increase in the market certainty-equivalent price p^* if the second good is more capital intensive than the

[9] The initial shareholders are indifferent to the purchase of the treasury shares, since the firm pays the market price. In this formulation the treasury shares are not utilized except to alter the cash flow per share to the remaining shareholders.

first. In contrast to Batra's result this holds for any strictly concave utility functions for consumers or for the managers in an entrepreneurial model and not just for the class of utility functions exhibiting decreasing absolute risk aversion as he finds. In contrast to Mayer's analysis the market certainty-equivalent is not stated in terms of a "change in expected price, with higher central moments constant" but instead is based on values that are readily observable in the securities market.

In a nonstochastic model, the Rybczynski theorem states that with a constant relative commodity price and inelastic factor supplies an increase in the supply of a factor increases the output of the commodity that uses that factor more intensively and reduces the output of the other commodity. Under uncertainty this theorem must be examined with a constant relative certainty-equivalent price, since as shown earlier, the presence of an international stock market guarantees that p^* will be the same in all countries. To analyze the effect of a change in the supply of K, it is first necessary to consider how the uncertain output of the first sector is to be treated. The output X_1 is given by $X_1 = \alpha L_1 f_1(k_1)$, so that an increase in $(L_1 f_1(k_1))$ due to an increase in K will result in an increase or decrease in actual output for any realization of α. The analysis will thus be made for the certainty portion $X_1^* = L_1 f_1(k_1)$ of output (and for $X_2 = L_2 f_2(k_2)$).[10] Differentiation yields

$$\frac{\partial X_1^*}{\partial K} = L_1 f_1'(k_1)\frac{\partial k_1}{\partial K} + f_1\frac{\partial L_1}{\partial K} \tag{26}$$

$$\frac{\partial X_2}{\partial K} = L_2 f_2'(k_2)\frac{\partial k_2}{\partial K} + f_2\frac{\partial L_2}{\partial K} \tag{27}$$

To determine the derivatives on the right sides of (26) and (27), totally differentiate (16), (17), (14), and (15), and set $dL = 0$ $(= dL_1 + dL_2)$ to obtain the following system of equations

$$\begin{pmatrix} k_1 - k_2 & L_1 & L_2 \\ 0 & p^* f_1'' & -f_2'' \\ 0 & -p^* k_1 f_1'' & k_2 f_2'' \end{pmatrix} \cdot \begin{pmatrix} dL_1 \\ dk_1 \\ dk_2 \end{pmatrix} = \begin{pmatrix} dK \\ 0 \\ 0 \end{pmatrix}$$

[10] Batra analyzes a change in the *ex post* output.

The determinant D of the coefficient matrix is $D = -p^* f_1'' f_2''(k_2 - k_1)^2 < 0$, and the solutions to the equations are

$$\frac{dL_1}{dK} = -\frac{dL_2}{dK} = \frac{1}{D}p^* f_1'' f_2''(k_2 - k_1)$$

$$= \frac{1}{k_2 - k_1}$$

$$\frac{dk_1}{dK} = \frac{dk_2}{dK} = 0$$

Evaluating the conditions in (26) and (27) yields

$$\frac{\partial X_1}{\partial K} = -\frac{f_1}{k_2 - k_1} \tag{28}$$

$$\frac{\partial X_2}{\partial K} = \frac{f_2}{k_2 - k_1} \tag{29}$$

If the first sector is more capital intensive than the second $(k_1 > k_2)$, then $\partial X_1/\partial K > 0$ and $\partial X_2/\partial K < 0$ which establishes the Rybczynski theorem. Technological and price uncertainty do not affect this basic result because the opportunity to trade in a securities market is sufficient to yield an observable market certainty-equivalent that may be used to plan input levels.

Batra and Das find, in contrast, that the Rybczynski theorem does not obtain because the relationship between the factor-price ratio and the commodity-price ratio depends in their model on the endowments in the economy. Mayer reaches a similar conclusion in his analysis of the intermediate run. In the model considered here the relationships given in (28) and (29) are independent of endowments, and, consequently, firms are able to plan their inputs using only information available in the securities and factor markets.

The Heckscher-Ohlin theorem follows directly from the Rybczynski theorem when the consumption patterns are identical in two countries, so a country exports the commodity that uses the more abundant factor more intensively. Also, the proof of the Stolper-Samuelson theorem is directly analogous to the proof in the certainty case. Differentiating (18) and the analogous condition for sector 2 yields

$$0 = f_1 - \frac{\partial w}{\partial p^*} - k_1 \frac{\partial r}{\partial p^*}$$

$$0 = -\frac{\partial w}{\partial p^*} - k_2 \frac{\partial r}{\partial p^*}$$

Solving gives

$$\frac{\partial r}{\partial p^*} = \frac{-f_1}{k_2 - k_1} \tag{30}$$

and

$$\frac{\partial w}{\partial p^*} = \frac{k_2 f_1}{k_2 - k_1}$$

Comparing (30) to (28) establishes the reciprocity theorem that

$$\frac{\partial X_1}{\partial K} = \frac{\partial r}{\partial p^*}$$

Similarly,

$$\frac{\partial X_1}{\partial L} = \frac{\partial w}{\partial p^*}$$

Converting (30) to elasticities and using (11) yields

$$\frac{dlogr}{dlogp^*} = -\frac{p^*}{r}\left(\frac{f_1}{k_2 - k_1}\right) = -\frac{w + rk_1}{r(k_2 - k_1)} = -\frac{\omega + k_1}{k_2 - k_1}$$

$$\frac{dlogw}{dlogp^*} = \frac{p^* k_2}{w}\left(\frac{f_1}{k_2 - k_1}\right) = \frac{k_2(w + rk_1)}{w(k_2 - k_1)} = \frac{k_2(\omega + k_1)}{\omega(k_2 - k_1)}$$

Consequently, if capital is used more intensively in sector 1 ($k_1 > k_2$), the real reward to capital increases with an increase in the relative certainty-equivalent price of the first commodity while the real reward to labor decreases. This establishes the Stopler-Samuelson theorem.

With irreversible factor intensities the factor-price equalization theorem will also hold as long as consumers in both countries can trade in the same securities and commodity markets. This is in contrast to Mayer's conclusion that the factor-price equalization theorem does not hold unless all "utility functions and probability assessments of a given industry's firms are identical in both countries" (p. 804).

IV. Conclusions

Although technological uncertainty will result in price uncertainty, the existence of an international securities market is sufficient to yield the standard theorems of international trade, since the opportunity to allocate risks in a securities market permits consumers to equate their marginal rates of substitution between shares and savings so as to establish a market certainty-equivalent that the firm may use to plan its inputs. By using the information implicit in stock market data in making their decisions, the firm will operate in the stockholders' interest, even when ownership is divorced from management. The key to this analysis is the existence of the market certainty-equivalent for the uncertain factors in the model. Given this and competitive behavior, the theorems of international trade result as in a deterministic model. When a market does not exist in which the certainty-equivalent can be established, the standard results will not in general obtain.

REFERENCES

K. J. Arrow, "The Role of Securities in the Optimal Allocation of Risk Bearing," *Rev. Econ. Stud.*, Apr. 1964, *31*, 91–6.

D. P. Baron, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.*, Oct. 1970, *11*, 463–80.

R. N. Batra, "Production Uncertainty and the Heckscher-Ohlin Theorem," *Rev. Econ. Stud.*, Apr. 1975, *42*, 259–68.

S. W. Das, "Uncertainty and the Heckscher-Ohlin Theorem: A Comment," *Rev. Econ. Stud.*, Feb. 1977, *44*, 189–90.

Gerard Debreu, *Theory of Value*, New York 1959.

P. Diamond, "The Role of the Stock Market in a General Equilibrium Model with Technological Uncertainty," *Amer. Econ. Rev.*, Sept. 1967, *57*, 759–76.

E. Helpman and A. Razin, (1976a) "International Trade Under Uncertainty: Recent Developments," work. paper no. 38-76, Foerder Inst. Econ. Res., Tel-Aviv Univ., June 1976.

——— and ———, (1976b) "Welfare Aspects of International Trade in Goods and Securities," work. paper no. 13-76, Foerder Inst. Econ. Res., Tel-Aviv Univ., Mar. 1976.

——— and ———, "Uncertainty and International Trade in the Presence of Stock Markets," *Rev. Econ. Stud.*, forthcoming.

Murray Kemp, *The Pure Theory of International Trade and Investment*, Englewood Cliffs 1969.

H. E. Leland, "The Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

W. Mayer, "The Rybczynski, Stolper-Samuelson, and Factor-Price Equalization Theorems under Price Uncertainty," *Amer. Econ. Rev.*, Dec. 1976, *66*, 797–808.

Multi-Intermediate-Goods Trade: The Gains and a Heckscher-Ohlin Analysis

By WILLIAM DER*

Three recent articles in this *Review* by Raveendra Batra and Francisco Casas (hereafter B-C), Albert Schweinberger, and James Riedel discussed the effects of intermediate goods on the factor proportions of trade à la Heckscher-Ohlin (hereafter H-O). Schweinberger (p. 634) interpreted the B-C paper as having introduced a nontraded intermediate input into the basic H-O model and having shown that the H-O model remains valid only if factor intensities are interpreted as direct-plus-indirect factor requirements (i.e., the total of all factor requirements at all stages of production). But a nontraded intermediate input runs contrary to B-C's only theorem dealing with the pattern of trade (Theorem 3); and in their entire section dealing with the pattern of trade they discuss the capital intensity of traded goods only in terms of k_i, their ratio of direct factor use (i.e., factor use only at the final stage of production). However, there is ambiguity in B-C's results since their equation (3) treats the intermediate good as being nontraded. Schweinberger's own treatment of prices as exogenous to the economy is contrary to the H-O model and permits each country to generally produce only two goods (which is the case Schweinberger examines) even if there were numerous goods.

Riedel (pp. 442–43, 446–47) asserts that Leontief's measurement of the direct-plus-indirect factor content of trade is an incorrect test of the H-O theorem because the input-output matrix assumes all intermediate products are produced domestically. In contradiction to Riedel, I will demonstrate that trade in intermediate goods is consistent with the input-output system and that the measurement of direct-plus-indirect factor content provides a test of the H-O theorem.[1] The preliminary portion of this paper will present the basic model and both a mathematical and a diagrammatical explanation of the gains from trade in intermediate goods for a small country.[2]

I. The Basic Model

The basic model assumes that there are two factors of production and any n number of goods (greater than one). It is assumed that intermediate goods are employed in fixed proportions while primary factors are employed in variable proportions. The assumption of fixed intermediate input coefficients permits use of the input-output system in an interindustry flows approach. The input-output system permits goods to be both intermediate and final products. The total domestic output of a good is referred to as *gross* output. The amount of the gross output of a good not used domestically as intermediate inputs is referred to as *net* output (or final output), which may be consumed domestically or exported. Furthermore, it is assumed that there are no joint products among the n goods.

Batra and Casas have asserted that their model differs from an interindustry flows (input-output) model in that the goods in their model are either final products or intermediate goods, but not both. Contrary to that assertion, the B-C model with its fixed intermediate input coefficients is just a special case of the present interindustry flows model, with appropriate net output variables and intermediate input coefficients set equal to zero.

*Iowa State University. This article is based on portions of my doctoral dissertation. I wish to thank George H. Borts, James A. Hanson, Kenneth S. Chan, Henry Ito, Arun Mukhopadhyay, and an anonymous referee.

[1]This supports a view taken by Robert Baldwin (pp. 132–33, fn. 11).

[2]See my dissertation, ch. 2, for an analysis of the gains for a larger country.

Batra and Casas also expressed the belief that "in a two-good model where intermediate and final products are identical, one cannot explain the basis of trade in intermediate goods" (p. 297). I will demonstrate that the basis for trade in intermediate goods can be explained for such a model. Throughout the analysis it will be assumed that the product and factor markets are perfectly competitive so that, in long-run equilibrium, product price is equal to product cost for those goods produced.

The following notation will be used:

X_i = gross output of good i, $(i = 1, \ldots, n)$
D_i = net output of good i, $(i = 1, \ldots, n)$
a_{ij} = the amount of good i required directly as an intermediate input in the production of one unit of good j, $(i, j = 1, \ldots, n)$

D and X will be column vectors whose components consist, respectively, of the D_i and the X_i, while $[a]$ is the $n \times n$ matrix of intermediate input coefficients. The input-output system for the closed economy is conventionally described by (1):

$$D = [I - a]X \tag{1}$$

For an existent inverse $A = [I - a]^{-1}$, X may be written as a linear function of D in (2):

$$X = AD \tag{2}$$

In order for system (2) to be solvable in nonnegative unknowns $X_i (i = 1, \ldots, n)$ for any vector of nonnegative D_i, a necessary and sufficient condition is that all principal minors of $[I - a]$ be positive. This is known as the Hawkins-Simon condition, and as is conventional, it is assumed that this condition is met. (See David Hawkins and Herbert Simon or Hukakane Nikaido, pp. 90–93.) The Hawkins-Simon condition implies the existence of $A = [I - a]^{-1}$ and the nonnegativity of the A_{ij} in system (2). For a closed economy, the A_{ij} would indicate the gross amount of good i needed in the productive activity of the entire system at all stages in order to yield one unit of net output of good j.

In addition, two primary factors, capital and labor, are assumed to be combined in variable proportions. The economy's endowment of capital and labor is fixed at the levels K and L, respectively.

Before international trade is introduced, the characteristics of the gross production possibility and net production possibility sets will be derived for the closed economy. The two-good case of the model will be discussed occasionally to facilitate the diagrammatics.

There are two constraints on the attainable levels of gross output for the closed economy. The first constraint is the scarcity of the primary factors of production. The second constraint is the requirement that the input-output production relations be satisfied; that is, the X_i and D_i must be nonnegative in satisfying (1) and (2).

The terms K_i and L_i denote the quantities of capital and labor, respectively, allocated directly to the production of good i. Thus, subject to the availability of intermediate inputs (which will be considered shortly), the gross output of commodity i is

$$X_i = F_i(K_i, L_i) \tag{3}$$

It is assumed that F_i is homogeneous of the first degree with diminishing, but positive, marginal physical products of the primary factors. Considering only the scarcity of primary factors, the functions in (3) generate a primary factor gross production possibility constraint, which is represented in gross output space by the curve $ABB'A'$ of Figure 1 for the two-good case. The term T_X denotes the set of points in gross output space which satisfy the primary factors gross production possibility constraint. T_X is a convex set (though not necessarily strictly convex) under the assumptions imposed upon the functions F_i and is represented for the two-good case by the points in the set $0ABB'A'0$ in Figure 1.

Consideration must still be given to the constraint upon gross output levels placed by the requirement for a closed economy that the input-output relations be satisfied wholly from domestic production. For instance, in Figure 1, point A is not feasible if the production of good 2 requires intermediate inputs of good 1, since at point A there would be no gross output of good 1 available for use as inputs if all primary factors were allocated directly to the production of good 2.

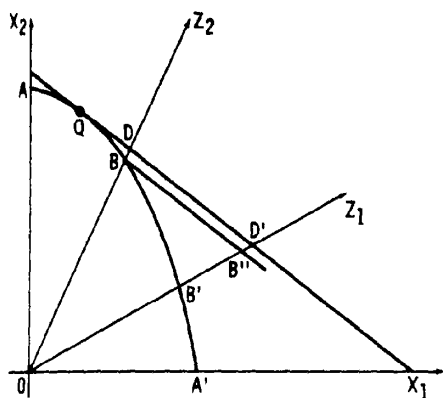



FIGURE 1

Because of its nonsingularity (i.e., its invertibility $[I - a]$ is a *one-to-one onto* mapping of all points in gross output space to all points in net output space.[3] That is to say, $[I - a]$ in (1) maps the set of all real valued vectors $(X_1, \ldots, X_n)$ one-to-one onto the set of all real valued vectors $(D_1, \ldots, D_n)$. The inverse matrix $A = [I - a]^{-1}$ is, of course, also nonsingular and maps together in (2) the same points as does $[I - a]$ in (1). A vector of gross output $(X_1, \ldots, X_n)$ is feasible with respect to input-output requirements in a closed economy only if it yields in (1) a nonnegative vector of net outputs $(D_1, \ldots, D_n)$. It is not feasible for a D_i to be negative since it would imply that the required level of good i for use as an intermediate input exceeds the gross output X_i. Thus, by the one-to-one *onto* mapping and the nonnegativity of the elements of A, a vector X of gross outputs satisfies the input-output requirements constraint if and only if it is a member of the set S.[4]

$$(4) \qquad S = [X \mid X = AD; \text{ for all } D \geq 0]$$

The convex set S of vectors satisfying the input-output requirements constitutes a polyhedral cone (See David Gale, p. 290.) Henceforth, the set S specified in (4) will be referred to as the *I-O polyhedral cone*. For the two-good case, the *I-O* polyhedral cone S is depicted in Figure 1 as the region bounded by and including the two rays OZ_1 and OZ_2. For the two-good case, the *I-O* polyhedral cone consists specifically of the vectors:

$$(5) \quad \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} D_1 \\ D_2 \end{pmatrix} = \begin{pmatrix} A_{11} \\ A_{21} \end{pmatrix} D_1 + \begin{pmatrix} A_{12} \\ A_{22} \end{pmatrix} D_2$$

for all $(D_1, D_2) \geq 0$. The rays OZ_1 and OZ_2 are, respectively, $X = (A_1 D_1 \mid D_1 \geq 0)$ and $X = (A_2 D_2 \mid D_2 \geq 0)$, where A_i is column i of the 2 x 2 matrix A in (5).

In Figure 1 the points contained within $OBB'O$ constitute the convex gross production possibility set for the closed economy since the points in this set satisfy both the scarce primary factors constraint and the input-output requirements constraint.[5] The set of points which compose the gross production possibility set for the closed economy (i.e., which are members of both T_X and S) will be called R_X.

The points in the gross production possibility set R_X are mapped one-to-one onto the convex net production possibility set (to be denoted by R_D) for the closed economy by the $[I - a]$ matrix.

$$(6) \quad R_D = [D \mid D = [I - a]X \text{ where } X \text{ is a member of } R_X]$$

The net production possibility set ($OEE'O$ in Figure 2 for the two-good case) consists of nonnegative points since they map from the points in R_X, which satisfy the input-output requirements for the closed economy. Points B and B' in Figure 1 map, respectively, to E and E' in Figure 2. The mapping is *onto* since the points in the net production possibility set are defined as the set of points in net output space mapped from the points in the gross production possibility set R_X. The convexity of the net production possibility set R_D follows

[3]A one-to-one mapping from a set A onto a set B is a mapping that identifies each element of A with exactly one element of B and each element of B with exactly one element of A.

[4]The condition $v' \geq v$ requires that each component of vector v' be equal to or greater than the corresponding component in vector v.

[5]The intersection of two convex sets is convex.

FIGURE 2

from the fact that R_X is convex and that linear transformations preserve convexity.

II. Small-Country Trade

The analysis of trade in intermediate goods will begin with the study of the gains from trade for a small country. The small-country assumption permits the country to export and import any quantities of goods at the exogenously determined vector of world prices with the restriction that the trade balance be zero.

The constraints under trade are the scarcity of primary factors constraint, the requirement that the balance of payments be zero, and the requirement that the levels of domestic gross production and imports satisfy at least the intermediate input needs of the domestic gross production levels.

The term M_i denotes the quantity of good i which is imported, with $M_i < 0$ if good i is exported; C_i will denote the domestic consumption level of good i. In the closed economy, D_i was used to denote both the net output and the consumption level of good i. But if net output is still considered to be the residual output remaining by subtracting domestic intermediate input needs from the gross output level (i.e., $D = [I - a]X$), then net output is not necessarily equivalent to the domestic consumption level since imports or exports will separate the two levels. The modified input-output system under trade is system (7) with the requirement that vector C be nonnegative.[6]

$$\text{(7)} \quad \begin{aligned} X_1 + M_1 &= a_{11}X_1 + a_{12}X_2 + C_1 \\ X_2 + M_2 &= a_{21}X_1 + a_{22}X_2 + C_2 \end{aligned}$$

Equation (7) says that domestic consumption C_i and domestic intermediate input use $a_{i1}X_1 + a_{i2}X_2$ of good i must equal the domestic gross production plus net imports $X_i + M_i$ of good i. The C_i are not permitted to be negative as $C_i < 0$ would imply that gross production and net imports of good i are not sufficient to meet intermediate input requirements of good i to produce the given levels of domestic gross output.[7]

It will be shown that for a small country with the opportunity to trade, any gross output vector X satisfying the scarce primary factors constraint (inside or outside of the *I-O* polyhedral cone) can satisfy the balance-of-payments constraint (8), where P_i is the world price of good i:

$$\text{(8)} \qquad \sum_{i=1}^{n} P_i M_i = 0$$

By multiplying each line i of (7) by price P_i and bringing to the left-hand side all terms except P_iC_i, (9) is obtained:

$$\text{(9)} \quad \begin{aligned} P_1M_1 + P_1X_1(1 - a_{11}) - P_1a_{12}X_2 &= P_1C_1 \\ P_2M_2 + P_2a_{21}X_1 + P_2(1 - a_{22})X_2 &= P_2C_2 \end{aligned}$$

[6] In this section analyzing trade, two-good forms of the matrices will be frequently used to facilitate the diagrammatical exposition. System (7) may be rewritten in general matrix form as $C = [I - a]X + M = D + M$.

[7] Riedel states that "the domestically produced inputs required directly and indirectly per unit of final demand" (p. 443), are given in a matrix $[I - a - m]^{-1}$, where $[m]$ is a matrix of per unit imported input coefficients. Using L to denote Riedel's vector of *domestically produced inputs* and F as the vector of final demand, what Riedel is saying is that $F = [I - a - m]L = L - aL - mL$. However, it is not apparent as to what use it is to define final demand F as L, the level of domestic production of intermediate goods, minus aL, the need for intermediate goods in producing the domestic production level of intermediate goods, and minus mL, the amount of intermediate inputs imported to produce the domestically produced intermediate inputs.

Adding the equations in (9) and factoring out the X_i's yields (10).

$$(10)\quad P_1M_1 + P_2M_2 + (P_1 - P_1a_{11} - P_2a_{21})X_1 + (P_2 - P_1a_{12} - P_2a_{22})X_2 = P_1C_1 + P_2C_2$$

Defining V_i as the unit price of product i minus the intermediate input costs directly incurred by a firm in producing a unit of product i implies

$$(11)\quad V_1 = P_1 - P_1a_{11} - P_2a_{21}$$
$$V_2 = P_2 - P_1a_{12} - P_2a_{22}$$

In matrix notation, (11) may be written as

$$(12)\quad V = [I - a']P$$

or as

$$V' = P'[I - a]$$

where V and P are column vectors; V_i is the *value-added* in producing a unit of good i. It is the price imputed to the activity of directly using labor and capital to produce a unit of good i. For a given vector of prices P, the firm knows V_i, the imputed price of its activity, if it knows the intermediate input coefficients.

The term X_i can represent the level of activity of the firms producing good i. The imputed value of the total level of activity i is V_iX_i. Since the level of national income Y, which is the summation of factor incomes, is equal to the summation of the values-added for the entire economy if w and r denote the factor prices of labor and capital, respectively; national income in equilibrium is

$$(13)\quad Y = wL + rK = V_1X_1 + V_2X_2$$

or

$$Y = V'X$$

By the definitions in (11), the V_i can be substituted into (10) to get (14).

$$(14)\quad P_1M_1 + P_2M_2 + V_1X_1 + V_2X_2 = P_1C_1 + P_2C_2$$

or

$$P'M + V'X = P'C$$

It is clear from (13) and (14) that as long as domestic expenditure on consumption equals the amount of income received, Y, at the given prices, the balance of trade will equal zero as required.

$$(15)\quad P'M = P'C - V'X = 0$$

For a chosen vector of gross output levels X, the national income is determined by (13). Equation (15) indicates that in satisfying the balance-of-payments constraint (8), the income receivers are free to purchase any vector of net domestic demands C, such that $P'C$, the value of net domestic demands, equals the national income. Since C is not uniquely related to X, there is no problem satisfying both $Y = P'C$ and $C \geq 0$ for any choice of X satisfying the scarce primary factors constraint.

There is no need to restrict production to the *I-O* polyhedral cone as was necessary for the closed economy. The gross production possibility set for the open economy is the set T_X (the set *0ABB'A'0* in Figure 1), which is larger than and contains the closed-economy gross production possibility set R_X. The vector of imports M satisfying the balance-of-trade constraint, the input-output requirements, and the final demands can be found by solving (7) once X and C are determined.

That the gross production possibility set expands to T_X when the economy is opened implies that the net production possibility set (composed of the points $D = [I - a]X$, where $X \in T_X$) expands beyond the set R_D. Since the small open economy is able to produce at gross output points outside of the *I-O* polyhedral cone, the additional net output points will have negative components. The term T_D will denote the net production possibility set for the open economy, where $[I - a]$ maps the points of T_X one-to-one onto T_D. The set T_D is convex as long as T_X is convex since linear transformations preserve convexity. In the two-good case of Figure 2, the net production possibility set T_D for the open economy is *0UEE'U'0*. Production outside of the non-negative orthant in net output space requires imports of intermediate goods to satisfy input-output requirements.

THEOREM 1: *D^* maximizes $P'D$ subject to $D \in T_D$ if and only if $X^* = [I - a]^{-1}D^*$ maximizes $V'X$ subject to $X \in T_X$.*

PROOF:

Suppose that $D^* \in T_D$ maximizes $P'D$ but that $X^* = [I - a]^{-1}D^*$ which is in T_X doesn't maximize $V'X$. Then some $\overline{X} \in T_X$ with $\overline{X} \neq$

X^* must maximize $V'X$. Since $\bar{X} \in T_X$, then $\bar{D} = [I - a]\bar{X}$ is in T_D. Now $P'D^* = P'[I - a]X^* = V'X^*$ and $V'\bar{X} = P'[I - a]\bar{X} = P'\bar{D}$. If $V'\bar{X} > V'X^*$, then $P'\bar{D} > P'D^*$, which contradicts the assumption that D^* maximizes $P'D$. Thus $X^* = [I - a]^{-1}D^*$ must maximize $V'X$ subject to $X \in T_X$.[8]

Since $Y = V'X$ is national income, then $P'D = P'[I - a]X = V'X$ is also national income. Along with the convexity conditions on the production possibility sets, it follows from Theorem 1 that the maximum national income can be found by finding the point or points of tangency of the $V'X$ budget hyperplane with the open-economy gross production possibility set T_X or, equivalently, by finding the point or points of tangency of the P^*D budget hyperplane with the open-economy net production possiblity set T_D.[9] The net output vector or vectors mapped by $[I - a]$ from the income-maximizing vector or vectors X^* are the same as the net output vector or vectors D^* which one would find by directly maximizing $P'D$ subject to $D \in T_D$. Marginal conditions may be expressed either in terms of V_i and X_i or in terms of P_i and D_i.

THEOREM 2: *If the gross output vector or vectors which maximize national income lie outside of the I-O polyhedral cone, then the open-economy consumption transformation set lies totally outside of the closed-economy consumption possibility set.*

The condition that the gross output vector or vectors which maximize national income lie outside of the *I-O* polyhedral cone is equivalent to the condition that the net output vector or vectors which maximize national income ($Y = P'D$) have at least one negative component. An example of the meaning of "totally outside" is the relation of the line *FF'* with respect to the closed-economy consumption possibility set *OEE'O* in Figure 2. However, a definition of totally outside is required.

DEFINITION: A set G lies totally outside of a set R_D 1) if for every element C^b in R_D, there exists a C^* in G such that $C^* > C^b$, and 2) if for every element C^* in G, there exists no C^b in R_D such that $C^b \geq C^*$.

The given conditions in Theorem 2 imply that none of the net output vectors in the closed-economy net production possibility set R_D maximizes national income when the economy is open. If the maximum national income is Y^* corresponding to some net output vector D^*, then the open economy's consumption transformation set is $G = [C|Y^* = P'C;\ C \geq 0]$. Since the closed economy's consumption possibility set is the same as its closed-economy net production possibility set R_D, it follows that for any consumption vector C^b in R_D, $Y^* > P'C^b$. Consider any n-component vector $e > 0$ satisfying $P'e = Y^* - P'C^b > 0$. Let $C^* = C^b + e$. Then C^* will be in the consumption transformation set G and $C^* > C^b$. This satisfies the first condition for totally outside.

To prove the second condition for totally outside, consider any vector C^* in G. Suppose that there exists a vector C^b in R_D such that $C^b \geq C^*$. This would imply that $P'C^b \geq P'C^* = Y^*$.[10] But this contradicts the condition that no vector C^b in R_D maximizes national income. Thus there will not exist a vector C^b in R_D such that $C^b \geq C^*$ if the conditions of Theorem 2 hold. This completes the proof of Theorem 2.

Theorem 2 shows that there is a potential gain from trade in intermediate goods. When the gross output vector or vectors which maximize national income lie outside of the *I-O* polyhedral cone (or equivalently, the net output vector lies outside of the nonnegative orthant and its bounding axes), the open-economy consumption transformation set lies

[8]The proof of the converse is left to the reader.

[9]It can be shown that all X vectors on the maximum $V'X$ budget hyperplane are mapped one-to-one onto the maximum $P'D$ budget hyperplane by $[I - a]$. See my dissertation, pp. 25–27. Thus if Q in Figure 1 is mapped to J in Figure 2 by $[I - a]$ then the budget line going through points Q, D, and D' in Figure 1 is mapped one-to-one onto the budget line going through points J, F, and F' in Figure 2.

[10]The inequality is true since the components of P are nonnegative.

totally outside of the closed-economy consumption possibility set, such as FF' lies totally outside of $OEE'O$ in Figure 2.[11] Without trade in intermediate goods, the open-economy consumption transformation set would be in the line EE' in Figure 2.

One might question whether it is theoretically possible to have world prices which would put the income-maximizing, gross output point or points outside of the *I-O* polyhedral cone (especially for the two-good case in response to Batra and Casas). Application of Rybczynski's theorem could move the income-maximizing gross output point or points outside of the *I-O* polyhedral cone if it is not outside already. Recall that the shape of the gross production-possibility frontier depends only on the linear homogeneous production functions in (3). The input-output coefficients determine the *I-O* polyhedral cone. Alteration of the factor endowment would move the income-maximizing point or points without changing the position of the *I-O* polyhedral cone.

III. A Heckscher-Ohlin Analysis

This section examines the factor proportions of trade between two countries with unequal factor-endowment ratios. In addition to the earlier production function assumptions it is assumed that production functions are identical across countries and that the direct factor-intensity rankings of the goods do not change for all wage-rent ratios. The countries are assumed to have identical homothetic tastes so that at the same relative prices, both countries desire to consume goods in the same proportion. The terms K^A and L^A denote the capital and labor endowments of country A, and K^B and L^B denote the capital and labor endowments of country B. It will be supposed that country A is relatively capital abundant compared to country B.

The objective of the analysis to follow is to demonstrate that the use of the input-output system, largely similar to Wassily Leontief's use (1953, 1956), to measure the direct-plus-indirect factor content of traded bundles in equilibrium is a correct test of a Heckscher-Ohlin-type model with trade in intermediate goods. No conclusion is drawn in this paper as to whether a measurement of only direct factor content is a correct test. From here on, unless otherwise stated, capital intensity or the capital-labor ratio is measured by direct-plus-indirect factor content. The analysis is divided into the two possible cases of factor-price equalization and unequal factor prices.

A. *Factor Prices Equalized*

The assumptions of perfect competition and positive marginal physical products in the functions of (3) imply that the primary factors will be fully employed. Idle resources can be allocated among the productive activities in proportions necessary to satisfy input-output requirements (even without imports) for additional output, with a possible shifting of fully employed resources (through factor substitution) to previously unproduced goods.

It will be recalled from (7) that the consumption vector C is equal to the sum of the net output vector $D = [I - a]X$ and the net import vector M:

$$C = D + M \tag{16}$$

Only vector C must have all nonnegative components. Any negative component of D indicates that the corresponding good is imported to satisfy the intermediate input needs of producing the vector D. Thus, this model encompasses the situations in which intermediate goods are traded.

Multiplying (16) by a row vector b of either all direct, all indirect, or all direct-plus-indirect capital or labor coefficients for the various goods yields

$$bC = b[D + M] = bD + bM \tag{17}$$

Equation (17) indicates that the factor content of C is equal to the factor content of D plus the factor content of M.[12] Attention here

[11] Ronald McKinnon and Robert Warne explain similar results in models with two goods.

[12] For example, if b is a vector of direct capital coefficients, then the first component of b indicates the direct capital required per unit of good 1, with the number of units of good 1 indicated in the vectors C, D, or M.

is directed only at the direct-plus-indirect factor content case.

To use Leontief's technique to calculate the amount of factors needed directly and indirectly to produce a net output vector D, one would premultiply the D by the $A = [I - a]^{-1}$ matrix to get the vector of gross outputs X which would satisfy gross output requirements at all stages of production for the production of the vector D. The amount of factors needed *directly* to produce the gross output vector X would constitute the direct-plus-indirect factors needed to produce vector D. Leontief's technique can also be used to compute the direct-plus-indirect factor requirements of consumption, exports, and imports.

Since the country's entire endowment of factors is exhausted directly in the production of the resulting gross output vector X, and since X results in the domestic net output vector D, this means that the direct-plus-indirect factor requirements for D equal the factor endowment of the country. Thus the direct-plus-indirect capital-labor ratio of each country's net output vector in equilibrium equals the country's capital-labor endowment ratio. Thus K^A/L^A will be the capital-labor ratio of A's net output, and K^B/L^B will be the capital-labor ratio of B's net output. Under factor-price equalization, goods would be produced with the same capital intensities across countries.

In the absence of tariffs, both countries will face the same relative prices in equilibrium. Consequently both countries consume goods in the same proportion because of identical tastes, and the capital-labor ratios of the two consumption bundles will be equal. Furthermore, the capital-labor ratio of each country's consumption bundle will equal the capital-labor endowment ratio for the world since all of the world's endowment of factors will be used directly and indirectly to produce the gross output necessary to yield the world's consumption bundle and since the world as a whole must satisfy the input-output requirements for the consumption vector.

Given that country A is relatively capital abundant, relation (18) holds:

$$\frac{K^A}{L^A} > \frac{K^B}{L^B} \tag{18}$$

Relation (18) implies[13]

$$\frac{K^A}{L^A} > \frac{K^A + K^B}{L^A + L^B} > \frac{K^B}{L^B} \tag{19}$$

where $K^W/L^W = (K^A + K^B)/(L^A + L^B)$ is the world capital-labor endowment ratio.

The terms E_K and E_L denote, respectively, the direct-plus-indirect capital and labor embodied in the exports of country A, using the Leontief technique of measurement; M_K and M_L denote, respectively, the direct-plus-indirect capital and labor embodied in imports of country A. (E_K, E_L, M_K, and M_L will be referred to as exports and imports of embodied capital and labor.) $\Delta K^A \equiv M_K - E_K$ will denote country A's net embodied capital inflow and $\Delta L^A \equiv M_L - E_L$ will denote country A's net embodied labor inflow.

Under the specified trade conditions, there must be a net flow of embodied factors so that each country is left with a consumption bundle of capital intensity equal to the world capital-labor endowment ratio. This is to say, condition (20) must hold in equilibrium:

$$\frac{K^A}{L^A} > \frac{K^A + \Delta K^A}{L^A + \Delta L^A} = \frac{K^W}{L^W} \tag{20}$$

where K^A is the capital embodied in D while ΔK^A is the capital embodied in the import vector M. By (17), the sum of K^A and ΔK^A is the capital content of C. When the capital content of C is divided by $L^A + \Delta L^A$, the labor content of C, then the ratio equals K^W/L^W as established earlier.

Since $L^A > 0$ and $L^A + \Delta L^A > 0$, multiplication of both sides of the inequality of (20) by L^A and $L^A + \Delta L^A$ yields

$$K^AL^A + K^A\Delta L^A > K^AL^A + L^A\Delta K^A \tag{21}$$

Subtracting K^AL^A from both sides of (21) and dividing by L^A implies

$$\frac{K^A}{L^A}\Delta L^A > \Delta K^A \tag{22}$$

[13]Equation (19) follows from $a/b > (a + c)/(b + d) > c/d$. The proof follows. Steps 2), 3), 4), 6), 7), and 8) are each implied by the step preceding it. 1) $a/b > c/d$; 2) $d > bc/a$; 3) $b + d > bc/a + ba/a = b(a + c)/a$; 4) $a/b > (a + c)/(b + d)$; and 5) $a/b > c/d$; 6) $ad/c > b$; 7) $ad/c + d = d(a + c)/c > b + d$; 8) $(a + c)/(b + d) > c/d$. Steps 4) and 8) together constitute the result sought.

Condition (22) implies three possible mathematical situations; however, only one situation is consistent with balanced trade required in the H-O model. The three mathematical cases are as follows:

Case 1: $\Delta L^A > 0$
which permits $\Delta K^A \gtreqless 0$ as long as $(K^A/L^A)\Delta L^A > \Delta K^A$.

Case 2: $\Delta L^A < 0$
which requires that $\Delta K^A < 0$ since $\Delta L^A < 0$ along with condition (22) implies that $0 > (K^A/L^A)\Delta L^A > \Delta K^A$.

Case 3: $\Delta L^A = 0$
which requires that $\Delta K^A < 0$ since $\Delta L^A = 0$ along with condition (22) implies $0 = (K^A/L^A)\Delta L^A > \Delta K^A$.

The assumption of perfect competition implies that product price will be equal to factor costs. Thus, the value of products traded is equal to the value of the factors embodied in those products traded. With factor-price equalization and identical production functions across countries, the movement of embodied factors is an indication of the value of trade occurring. The value in terms of a numeraire of the factors embodied directly-plus-indirectly in the traded bundles must equal the value in terms of the numeraire of the market value of the traded bundles. The cases in which net inflows or outflows of embodied factors imply that trade is not balanced can be disregarded as they would not be consistent with the balanced trade assumption implicit in the H-O model equilibrium.

Cases 2 and 3 can be ruled out as equilibrium situations as they would necessarily involve trade imbalances. Case 2, for instance, is one in which there is a net outflow of both embodied labor and embodied capital. This case implies that country A's exports contain more embodied capital and embodied labor than its imports. Consequently, the value of A's exports must exceed the value of its imports. In this situation, there would be a trade surplus for country A. Case 3, for similar reasons, also implies a trade imbalance. If for a country there is a net export of one embodied factor, then trade balance requires that there be a net import of the other embodied factor. Only in the subcase of Case 1 in which $\Delta L^A > 0$ and $\Delta K^A < 0$ can trade possibly be balanced. Thus the necessary condition for trade balance in the H-O model is that country A be a net importer of embodied labor and a net exporter of embodied capital. This requires that

(23) $$\Delta L^A = M_L - E_L > 0$$
and $$\Delta K^A = M_K - E_K < 0$$

or, equivalently,

(24) $$M_L > E_L \quad \text{and} \quad M_K < E_K$$

Dividing both sides of the second inequality in (24) by M_L yields

(25) $$\frac{M_K}{M_L} < \frac{E_K}{M_L}$$

Since $M_L > E_L$, then condition (26) holds:

(26) $$\frac{E_K}{M_L} < \frac{E_K}{E_L}$$

Conditions (25) and (26) together imply the H-O conclusion:

(27) $$\frac{M_K}{M_L} < \frac{E_K}{E_L}$$

This completes the proof that if factor-price equalization occurs the exports of the capital abundant country would be capital intensive relative to its imports when capital intensity is measured by direct-plus-indirect factor content using the input-output system.

B. *Unequal Factor Prices*

Attention is again restricted to trade without tariffs. The absence of tariffs compels the analyst to rely upon foreign factor coefficients for goods not produced domestically in order to measure the factor content of trade. But this is a predicament an analyst would face in testing a factor-proportions theory of trade even if only direct factor coefficients were measured or if there were no intermediate goods.

It will be shown that with unequal factor

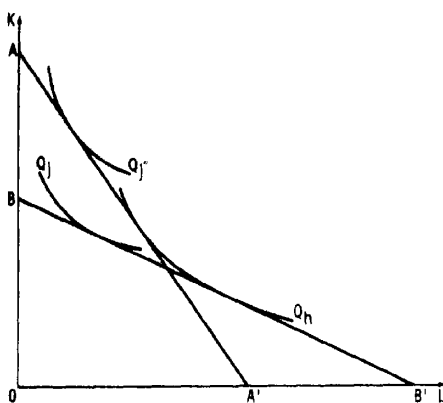



FIGURE 3

prices, the country (say, country I) with the higher wage-rent ratio will produce a chain of the most capital-intensive goods measured on only a direct factor basis, while the country (say, country II) with the lower wage-rent ratio will produce a chain of the least capital-intensive goods measured on a direct factor basis. Without tariffs, the countries will produce, at most, one good in common. These hypotheses will be demonstrated with the aid of Figure 3.

Given that both countries can purchase intermediate inputs at the same relative prices in the absence of tariffs, the world pattern of production will depend upon which country can produce which commodities at the lowest relative cost of direct factor use. In comparing countries, the direct factor cost can be measured in terms of the amounts of other goods of which a country must reduce production or in terms of the relative amount of expenditure it must add to the cost of intermediate inputs to produce the good.

Suppose that the goods are ordered in terms of increasing direct factor capital intensity and that both countries produce good h in equilibrium. In Figure 3, two factor-budget lines are drawn tangent to the same unit isoquant of direct factor use at the final stage of production for good h. Factor-budget line AA' represents a higher wage-rent ratio than line BB'. Since the points along any one budget line represent equal levels of expenditure, for the same level of expenditure country I could produce either one unit of good h or Q_j^* amount of any other good j, which has a higher direct factor capital intensity than good h. On the other hand, country II can produce either a unit of good h or Q_j amount of good j for a given level of expenditure in its own country. Since $Q_j^* > Q_j$ (indicating a lower cost relative to good h of producing good j in country I), since both countries can purchase intermediate goods at the same relative prices, and since product price equals product cost in equilibrium, it follows that if country I in equilibrium produces some good h, then it will alone be the country to produce each good with a direct factor capital intensity greater than good h since it can do so at a lower cost relative to the cost of good h than can country II. The same line of reasoning will establish that country II will alone produce goods with lower direct factor capital intensities than good h.

With unequal factor prices, the countries would not produce more than one good in common. The common production of two goods would generally require that the factor-budget lines for both countries have equal slopes, which violates the condition of unequal factor prices.

It may turn out that the two countries produce no goods in common. Suppose that one of the goods produced by country II is good g and that one of the goods produced by country I is good j. Suppose that in terms of a numeraire the value of the final stage of production (i.e., value-added) of a unit of good j in country I is equivalent to t units of value-added for good g in country II. If one draws a factor-budget line reflecting country I's higher wage-rent ratio tangent to the unit direct factors isoquant for good j and draws a factor budget line reflecting country II's wage-rent ratio tangent to the direct factor iosquant for t units of good g, one can use the same line of reasoning which accompanied Figure 3 to show in terms of direct factor content, good j must have a higher capital intensity than good g and that country I produces a chain of the most capital-intensive goods while country II produces a chain of the least capital-intensive goods.

It will turn out that country I is country A, the capital-abundant country. If that were not true then country A's gross output would have

a lower direct factor capital-labor ratio than country B's gross output. If country B is fully employed then the direct factor capital-labor ratio of B's gross output must equal its capital-labor endowment ratio. But that would imply that country A's factor-endowment ratio is greater than the direct factor capital-labor ratio of A's gross output. But this would contradict the condition of full employment which would occur with the assumption of perfect competition and positive marginal products. Thus, to avoid the contradiction, the capital-abundant country must produce the chain of the most capital-intensive goods in terms of direct factor use. Because of the described patterns of production, country A must in terms of direct factor content export only the most capital-intensive goods while importing only the least capital-intensive good.

When the factor content of trade is measured by direct-plus-indirect factor content using the input-output system, where foreign factor coefficients are used for intermediate goods which are not produced domestically, the capital-labor ratio of A's exports will also be greater than the capital-labor ratio of its imports. Suppose that E is country A's vector of exports with zeros in the arguments for goods not exported by A. When E is premultiplied by $[I - a]^{-1}$, the result is X_E, the vector of gross outputs necessary in total to yield the export bundle E. The direct-plus-indirect factor content of E is determined by summing the direct factor requirements necessary to produce the X_E vector of gross outputs. Those components of X_E which denote the foreign-produced intermediate inputs would tend to make the direct-plus-indirect capital intensity of E lower than the direct capital-intensity of E. However the direct and indirect source of these foreign-produced intermediate inputs is A's imports. Thus the portion of the direct-plus-indirect factor content of E accounted for by foreign-produced intermediate inputs has its equivalent in the direct-plus-indirect factor content of country A's imports. Similarly for B, a certain portion (relating to intermediate inputs originating in A) of the direct-plus-indirect factor content of country B's exports has its equivalent in the direct-plus-indirect factor content of country B's imports. The remaining portion of the direct-plus-indirect factor content of each country's exports is accounted for by the direct factor content of domestically produced goods. Because of the production chains established earlier, the direct-plus-indirect capital-labor ratio of country A's exports exceeds that of its imports.

IV. Concluding Remarks

An imposition of tariffs may be a way to avoid the reliance on foreign factor coefficients for testing a factor-proportions theory of trade. But complications will arise in the derivation of a theory. For one thing, the differences in product prices between the two countries may require more specific conditions on consumption patterns. For another thing, it may not be possible to use chains of goods to determine which country produces or exports the more capital-intensive bundle of goods. Furthermore, it may not be a simple exercise to show which country will have a higher wage-rent ratio (if it is still relevant) since the tariffs may be superimposed upon a structure which would have factor-price equalization without tariffs.

REFERENCES

R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, *61*, 126–46.

R. N. Batra and F. R. Casas, "Intermediate Products and the Pure Theory of International Trade: A Neo-Heckscher-Ohlin Framework," *Amer. Econ. Rev.*, June 1973, *63*, 297–311.

J. N. Bhagwati, "The Heckscher-Ohlin Theorem in the Multi-Commodity Case," *J. Polit. Econ.*, Sept./Oct. 1972, *80*, 1052–55.

Richard E. Caves, *Trade and Economic Structure*, Cambridge, Mass. 1960.

W. Der, "Aspects of Trade in Many Goods and in Intermediate Products," unpublished doctoral dissertation, Brown Univ. 1976.

D. Gale, "Convex Polyhedral Cones and

Linear Inequalities," in Tjalling C. Koopmans, ed., *Activity Analysis of Production and Allocation*, New York 1951.

D. Hawkins and H. A. Simon, "Note: Some Conditions of Macroeconomic Stability," *Econometrica*, July/Oct. 1949, *17*, 245–8.

W. W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Amer. Phil. Soc. Proc.*, Sept. 1953, *97*, 332–49.

——, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, *38*, 386–407.

R. I. McKinnon, "Intermediate Products and Differential Tariffs: A Generalization of Lerner's Symmetry Theorem," *Quart. J. Econ.*, Nov. 1966, *80*, 584–615.

Hukakane Nikaido, *Convex Structures and Economic Theory*, New York 1968.

J. Riedel, "Intermediate Products and the Theory of International Trade: A Generalization of the Pure Intermediate Good Case," *Amer. Econ. Rev.*, June 1976, *66*, 441–47.

P. A. Samuelson, "The Gains from International Trade Once Again," *Econ. J.*, Dec. 1962, *72*, 820–29.

A. Schweinberger, "Pure Traded Intermediate Products and the Heckscher-Ohlin Theorem," *Amer. Econ. Rev.*, Sept. 1975, *65*, 634–43.

J. Vanek, "Variable Factor Proportions and Inter-Industry Flows in the Theory of International Trade," *Quart. J. Econ.*, Feb. 1963, *77*, 129–42.

R. D. Warne, "Intermediate Goods in International Trade with Variable Proportions and Two Primary Inputs," *Quart. J. Econ.*, May 1971, *85*, 225–36.

David M. Winch, *Analytical Welfare Economics*, Baltimore 1971.

Information, Entry, and Welfare: *The Case for Collusion*

By DONALD DEWEY*

In the last thirty years many of the propositions about the "monopoly problem" which are used to justify an antitrust policy have been called into question and, in many instances, conclusively disproved (see, for example, Robert Bork, Harvey Goldschmid et al., and Richard Posner). As the intellectual foundations of antitrust have crumbled and collapsed, one has so far retained both its nearly universal acceptance and its academic respectability. Almost without exception, economists and lawyers have continued to assume that the case against price fixing—collusion—is so self-evident that it does not require detailed examination. Indeed, the hostility to price fixing is the one feature of American antitrust that seems to be exportable—witness, for example, its continuing incorporation into the cartel policy of the European Common Market. The purpose of this article is to suggest that even the "evil" of collusion can no longer be taken for granted—and, by implication, that a complete and skeptical review of the conspiracy and tacit conspiracy doctrines of antitrust is called for.

One might claim a certain urgency for this task. In recent years both the Congress and the antitrust agencies have given a very high priority in antitrust enforcement to increasing the number of prosecutions for price fixing and information sharing. If we are going down the wrong road, we are traveling at an increasing rate.

In retrospect, it is quite surprising that economists have allowed the case against collusion to go unexamined for so long. After all, the textbook treatment of collusion has always been tautological. It assumes that the object of price fixing is to increase the rate of return on capital; and that, if entry into the industry is free, price fixing cannot achieve this goal "in the long run." Therefore, so the textbooks imply, if a cartel endures over time, entry is not free and price fixing must have achieved its goal of creating a monopoly rent.

One can, of course, argue that price fixing may be forever attempted—and forever frustrated—when entry into the industry is free because businessmen are so stupid that they do not learn from experience. Such an assumption is, for good reason, usually rejected by economists. In any event, if it could be substantiated, a legal rule against price fixing would have to be defended on the strange ground that it protected the public against the folly of businessmen in squandering resources on a vain quest.

The textbook treatment of collusion presumes the existence of barriers to entry and the possibility of monopoly rent. But casual observation shows that in the absence of antitrust harassments price fixing will be found, most of the time, in industries where by any reasonable use of language entry is free and no obvious monopoly rents are being collected. Every industry that merits the name has at least one trade association. Inside every competently run trade association is a cartel yearning to breathe free of legal restraints.

Nor do we have to rely on casual observation for evidence that collusion will persist even when it does not lead to exceptional profitability. In recent studies (1975, 1976), Peter Asch and Joseph Seneca found that industries whose member firms have been prosecuted for price fixing by the antitrust agencies actually earned a below-average rate of return on capital. (As one would expect, these industries contain many small firms and entry and exit are easy.)

It is the thesis of this article that the

*Professor of economics, Columbia University. In the preparation of this article I have imposed upon the good nature of many critics. My debts are especially heavy to Joseph Seneca, W. G. Shepherd, and William Vickrey.

paradox of persisting efforts at price fixing in industries where entry is free and the rate of return on capital "low" has a plausible explanation; and, indeed, that the explanation has often been advanced by businessmen in their own inarticulate and self-serving way. When businessmen seek to justify price fixing, they invariably do so in terms of "preventing ruinous competition," "achieving a healthy climate for investment," "promoting stable market conditions," or some such reassuring rhetoric. Impatient economists usually translate the businessman's inept defense of price fixing as a simple desire to maximize profits.[1]

Let us try a different translation. Let us assume that what the businessman is trying to maximize is a utility function with two elements: the expected rate of return on capital and the variance associated with this rate.[2] (I introduce variance into the businessman's utility function simply because it is the most widely used measure of uncertainty; any other statistical measure of uncertainty would serve the purposes of this paper just as well.) In a world of uncertainty, this assumption is obviously more realistic than a premise of simple profit maximization. More important, we shall find that it is more useful for understanding the causes and consequences of collusion.

Given free entry, collusion cannot create a permanent monopoly rent. So much is certain. But it is not immediately apparent how collusion will affect the tradeoff between the rate of return on capital (henceforth called expected profit) and profit variance. For example, for given utility level, *if* collusion reduces profit variance it must of course reduce expected profit. Nor is it immediately apparent how collusion affects industry output in a free entry situation.

Given its popular connotations of evil purpose, secret maneuvering, and general wrongdoing, the term collusion is not well suited to the vocabulary of technical economics. What this paper means by collusion is better conveyed by the pedestrian phrase, "cooperative action that affects prices." However, I will continue to speak of collusion because of the term's popularity and brevity. By my usage, the term covers both agreements which directly affect price and output and what the British call information agreements, for example, agreements to share price and production data.

I. Conventional Assumptions

ASSUMPTION 1: *Production takes place through successive periods in an industry where uncertainty cannot be reduced below some minimum which is constant; that is, uncertainty can never be completely eliminated through "learning by doing" even though firms are allowed to exchange price and output data and negotiate agreements governing price and output.*

ASSUMPTION 2: *The industry consists of firms which sell a homogeneous product, not in a perfect market, but in one made imperfect by the costs of acquiring information—a Marshallian market for short. (I implicitly use the definition taken from formal information theory which makes information "that which reduces uncertainty.")*

ASSUMPTION 3: *Entry is free in the usual sense of the phrase: The entry of newcomers is not impeded by legislation, ignorance, an imperfect capital market, or technical handicap.*[3]

ASSUMPTION 4: *All firms are risk averse and seek to maximize the utility function $U(r,v)$ where r is expected profit and v is profit variance. Thus if $r_1 = r_2$ and $v_1 < v_2$, then $U(r_1,v_1) > U(r_2,v_2)$ while if $v_1 = v_2$ and $r_1 < r_2$, then $U(r_1,v_1) < U(r_2,v_2)$.*

[1]In all fairness, it should be mentioned that economists who have studied particular industries (especially so-called distressed industries) have sometimes given a sympathetic hearing to the businessman's defense of collusion. See, for example, the treatment of shipping cartels by Daniel Marx; see also George B. Richardson.

[2]Suppose that in period t the rates of return on capital, $r_1, r_2, \ldots, r_p$ are earned by P firms and that r^* designates the mean rate of return. To simplify the exposition, we shall assume that in period t + 1, expected profit is r and profit variance v is $\Sigma(r^* - r_i)^2/P$.

[3]Free entry, however, is compatible with the existence of a fixed cost in the firm even though the magnitude of this cost is one of the determinants of the equilibrium number of firms. See my paper.

ASSUMPTION 5: *All firms are replicates of one another; any proposition about one firm is a proposition about every other firm.*[4]

ASSUMPTION 6: *Let k denote the combination of r and v and E(K) the set of k's consistent with a size population of firms that is constant over time. Then our industry is in long-run equilibrium if, and only if,* $k \epsilon E(K)$. *Every k included in E(K) must, of course, have the same utility rank. In this (Marshallian) equilibrium, ex post, different firms will earn different profits and some firms may exit and be replaced by an equal number of others.*

II. Additional Assumptions

The assumptions set down so far would seem to require no elaboration. They have been so often explicitly employed and defended by economists that their usefulness can be presumed. The remaining assumptions deserve some words of explanation. For while they may well be as widely used as the above assumptions (especially in theorizing about "workable" competition or, more recently, in theorizing about the search process in markets), they are seldom formally stated.

The first of the remaining assumptions is made necessary by the tenacious hold of static price theory on all of us. In the usual textbook model, with its perfect market, the power of the firm to affect price is equated with "monopoly power." In this model, with its Marshallian market, the power of the firm to affect price is traceable to the presence of information costs in the market. This power is a short-run phenomenon in the sense that it would disappear if learning by doing could ever reduce search cost to zero. By Assumption 1 this cannot happen. Uncertainty—or entropy to use the term borrowed by formal information theory from statistical mechanics—will not fall below some minimum. Hence, at any moment, the firm in a Marshallian market always possesses the limited power over price that exists because information about prices is not free.[5]

Here we meet a predictable difficulty. When sellers are few in a Marshallian market, the firm may have both types of power—and disentangling their effects is not easy. One complication could be especially troublesome—the well-known result that, when entry is free and firms are few, each firm may produce an output at which unit cost is falling.

In this paper our concern is with the limited power over price traceable to an imperfect market—not with the power over price traceable to imperfect competition in a perfect market. Therefore, let us get rid of the complications of monopoly power by making the equilibrium number of firms great enough to insure that the single firm believes that, whatever its short-run power to influence price, its unilateral actions cannot affect the *average* of prices prevailing in the production period. (Just as the used car dealer can take pride in his skill in bargaining in particular transactions and still believe that it has no perceptible effect on the average of used car prices in his market for the year.) We eliminate the effect of monopoly power due to fewness of sellers by assuming:

ASSUMPTION 7: *Let x^* be the output at which unit cost is minimized in the firm. Then whatever the legal rules regarding collusion, in equilibrium the firm's output is expected to fluctuate during the production period about a mean of x^*. (This assumption does not exclude the possibility that the unit*

[4]G. Warren Nutter and John H. Moore have recently argued that competitive behavior which they equate with a firm's willingness to make unilateral price cuts (their analysis does not cover unilateral price rises) depends upon differences in the taste for risk among sellers. Such differences clearly place a constraint upon the amount of collusion that will be attempted in a market; in this sense they promote competitive behavior. But they are not a necessary condition for it. So long as collusion is not costless—and it seldom is—firms will invest in collusion up to the level where its expected marginal benefit equals its expected marginal cost. Under conditions of enduring uncertainty, rival firms, whatever their respective tastes for risk, must expect to make a number of price changes each period. It will ordinarily pay them to coordinate some—but not all—of these adjustments. Hence, there is no good reason to forego the simplification in exposition that Assumption 5 makes possible. (All firms have the same taste for risk.)

[5]On the search process in what I have called a Marshallian market, see George J. Stigler, and Nutter and Moore.

cost of x^* is affected by the legal rules regarding collusion.)

The next assumption requires an even more extended introduction. In the perfect market ruled by the Law of One Price, there can be no such thing as price competition. This gives us one more reason for distrusting generalizations about the effects of collusion based upon economic models which assume a perfect market. In our Marshallian market made imperfect by the costs of acquiring information, price competition does exist. It has two features especially relevant to this analysis.

The price changes which firms make during the production period are imperfectly coordinated. Each firm associates more than one possible payoff with every unilateral price change and, for each firm, the pricing behavior of rivals is a source of uncertainty. In general, the greater the number of price changes that they make, the greater this uncertainty.

In a Marshallian market, price changes by the firm serve two related but distinguishable purposes. They are the economic application of information already received. And they are a means of acquiring additional information since, in an uncertain world, some part of the activity of the firm must always involve an exploratory search for information.

Strictly speaking, there can never be price stability in the sense of unchanging prices in a Marshallian market. Firms will, in every period, vary prices to obtain information and the value of any bit of information once obtained will decline with the passing of time.[6]

In this analysis, we shall pass to the limit in simplification and use as our index of price competition the total number of price changes made by all firms in the industry during the production period. N will denote this number. Price competition has, of course, more dimensions than the total number of price changes per period; they include, at the very least, the presence or absence of a discernable pattern of price leadership in the price changes. But little will be lost by equating price competition with the size of N. It is virtually impossible to conceive of any change in price competition that does not involve a change in N. Looking to future empirical work, another merit may be claimed for N as an index of price competition. It can be directly estimated from price data.

Also, the use of the total number of price changes per period as an index of price competition obviates the need for any recondite speculation about how price competition is to be distinguished from collusion. By the usage of this paper, collusion is simply the absence of price competition. In this model, with its Marshallian market, anything that increases N increases price competition and simultaneously reduces collusion, and vice versa.

Nor do we need to get bogged down in efforts to classify the varieties of collusion. It is collusion when firms directly reduce N by agreeing to charge a common price and alter it only when acting together. It is collusion when firms pool their data on prices and costs in order to gain a more accurate picture of "underlying business conditions" and the result is a reduction in N. Collusion may be overt as in the formal cartel agreement or implicit as in the mental telepathy that produces price leadership in highly concentrated industries.

I now state the final three assumptions.

ASSUMPTION 8: *Let the legal system be given. Then for every industry output (hereafter designated X) there is a number of price changes which will maximize $U(k)$ for that output. If this number is exceeded,* (i) *r will fall,* (ii) *v will rise, or* (iii) *r will fall and v will rise.*

ASSUMPTION 9: *Let the legal system be given. Then the number of price changes which maximizes $U(k)$ for each output varies directly with output.*

ASSUMPTION 10: *The legal system matters. A rule change which reduces the cost of collusion will, by encouraging agreements to reduce N, have some effect upon r, v, and X.*

[6]Not many efforts have been made to apply formal information theory to economic problems. See, however, Henri Theil and R. A. Jenner.

Let us pause to emphasize one implication of the last three assumptions. Suppose that within a legal system that prohibits collusion, $U(k)$ for an industry could be raised if this prohibition could be ignored with impunity. Suppose also that such is not the case—that the legal system has sufficient teeth (in terms of the severity and probability of punishment) to deter collusion, that is, joint action to reduce N. Then by our usage, the N that actually results is "optimal." In short, to the extent that legal rules constrain business behavior, the N that is optimal from the industry's standpoint is partly a function of the legal system.

FIGURE 1

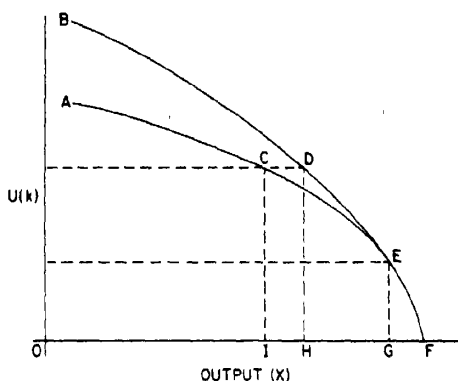
III. Alternative Legal Systems

We can now consider how the operation of our industry is affected by the law's treatment of collusion. Let us examine three possible legal systems.

(1) *Alpha System*: The industry, in common with all other industries, is subject to a legal code which prohibits, and effectively prevents, collusion. Defendants are ferreted out by an army of informers, tried by courts-martial, enjoy only the due process accorded to privates in the armies of Old Prussia, and, if convicted, sentenced to a reopened Alcatraz.

(2) *Beta System*: The industry is given a total exemption from the above Draconian version of antitrust while all other industries remain subject to it. To avoid ambiguity, let us make the price and information agreements of the exempted industry legally enforceable though, so far as economic effects go, it probably makes little difference whether the exempted industry uses the courts or private cartel machinery to enforce such agreements.

(3) *Omega System*: There is no antitrust for anyone. In all industries, firms have complete freedom to exchange information and negotiate price agreements. Otherwise Assumptions 1–10 remain in force.

By Assumption 10, "the legal system matters." But let us be more specific. 1) How will the granting of an antitrust exemption to a single industry (shifting from the *Alpha System* to the *Beta System*) affect its equilibrium output? 2) How will the granting of this exemption affect the equilibrium values of r and v? We know, of course, that $U(k)$—the utility rank of the $\{r,v\}$ combination—will be the same in both equilibria since all equilibrium k's must, by definition, have the same rank.

Let us begin by noting an "obvious" truth of crucial importance. For our chosen industry, the set of production possibilities represented by the *Beta System* of laws completely dominates that of the *Alpha System*. Any contract that can be negotiated under the *Alpha System* can also be negotiated under the *Beta System*, but the converse is not true. Under both legal systems the industry can refrain from colluding. Only under the *Beta System* is collusion feasible. The implications of this domination are indicated by Figure 1.

Here the curve *BDEF* gives the maximum utility rank of the k corresponding to each industry output when produced under the *Beta System* (our industry alone exempt from antitrust). The curve *ACEF* gives the maximum utility rank of the k corresponding to each industry output when produced under the *Alpha System* (antitrust for everyone).

Each curve in Figure 1 is drawn on the assumption that N has been optimized for each X under the legal system that the curve represents. In accordance with Assumption 9, movement along either curve rightward from the origin implies an increase in N. Likewise, in accordance with Assumption 10, whenever the two curves do not coincide, any given X

implies a lower N if produced under the *Beta System* than if produced under the *Alpha System*. When the two curves coincide in Figure 1, a given X will be associated with the same N under both systems; this is the case where law does not matter.

Suppose that the utility rank of every k consistent with equilibrium output is given by the vertical distance EG. Then equilibrium output is given by the horizontal distance $0G$; and $0G$ will be the industry's equilibrium output whether it is produced under the *Alpha* or *Beta System*. I have constructed Figure 1 to incorporate the popular—and perfectly reasonable view—that when the number of firms is "very large" it will not pay an industry to attempt collusion even when it is lawful (by Assumption 7, in the model the number of firms varies directly with industry output). With very large numbers, the contract costs of collusion are simply "too high."

Suppose, however, that the utility rank of every k consistent with industry equilibrium is given by the vertical distance HD (or IC). Now equilibrium output is $0I$ under the *Alpha System* but the greater quantity $0H$ under the *Beta System*.

We have answered our first question. Legalizing collusion in a single industry, when entry is free, will either leave equilibrium output unchanged or cause it to increase. Legalizing collusion will cause equilibrium output to increase when it operates to increase the number of contracts that it is economically rational to negotiate. This conclusion is, of course, perfectly compatible with the possibility—indeed the virtual certainty—that the immediate short-run consequence of legalizing collusion will be a contraction of industry output to the extent entry is slow to gain a higher r and lower v. Presumably movement to a new equilibrium will take time.

What of our second question—the effect of legalizing collusion on profit and profit variance? Here Figure 1 yields an ambiguous and intriguing result. Each of the following outcomes is possible in the new equilibrium: (i) r and v are unchanged, (ii) r and v are both greater, and (iii) r and v are both lower. The only requirement is that, in the new equilibrium, the combination $\{r,v\}$ has the same utility rank as that consistent with equilibrium in the *Alpha System*.

Why the ambiguity? Surely our initial expectation is that, if collusion is a way of reducing uncertainty, legalizing collusion in an industry where entry is free ought to produce a new equilibrium characterized by lower r and lower v. The pitfall in this expectation becomes apparent when we recall that the number of price changes per period is a function of both the legal system and the industry's output; and that legalizing collusion can lead to greater output.

Suppose that legalizing collusion does lead to greater output. Then such a change in the law has reduced the optimal N asssociated with each X over some range. (This range is given by the distance $0G$ in Figure 1.) However, as the industry moves to a new equilibrium, the increase in X produces an increase in N. Depending upon the strength of the opposing forces operating on N in the new equilibrium, legalizing collusion will leave r and v unchanged, lower both r and v, or raise both r and v. I can only say the obvious: the greater the increase in output that occurs when an industry is permitted to collude, the more likely that collusion will actually result in higher equilibrium values for both r and v.

To anyone with a traditional training in economics the conclusion that collusion can increase output is unfamiliar and possibly suspect. Nevertheless, to avoid it one would have to show that the statement that "collusion is effective" is a logical contradiction of one or more of the assumptions that we have employed. So long as collusion confers no benefits on established firms that are not available to new firms, no such demonstration can be made. But if information made possible by collusion is available only to established firms, Assumption 3 is contradicted; and they gain a cost advantage over potential rivals that would not exist in the absence of collusion.[7] (In this case collusion may—or

[7]An apparent instance of the use of information by insiders to attempt to block the entry of a new firm into the market is described in *Paterson Parchment Paper Co v. Story Parchment Co.* In this case the country's three established producers of parchment paper (then used to wrap meat) refused to accept a newcomer into their

may not—be output-increasing, i.e., collusion may be justified on welfare grounds even though it confers a cost advantage on established firms.) Any possibility that established firms may gain a competitive edge through inside information can, of course, be removed by making open membership a requirement for permitting collusion.

IV. When Collusion is Everywhere Legal

When only one industry receives an exemption from an otherwise universally enforced legal rule against collusion, the exempted industry can expand output mainly by attracting resources from other industries. If a policy condemning collusion is scrapped for all industries, we cannot say how any particular industry would be affected by this change without further information. The only safe generalization is that resources will move from industries where the rewards from the newly legalized collusion are low to those where they are high.

For the economy as a whole, the effects of legalizing collusion are more easily seen. Collusion will reduce uncertainty and therefore the cost of investment and marketing mistakes. In a risk-averse world, collusion by reducing uncertainty will presumably also increase the fraction of national income invested.

V. Collusion and Economic Welfare

So far we have stayed strictly within the confines of positive economics. The time has come to be more adventurous and we need have no hesitation in making the following generalization: In a model incorporating the assumptions used in this analysis (most notably the assumption of free entry) economic welfare will be greater under a legal system that permits collusion than under one which effectively suppresses or restricts it. Still, every economic model has its limitations as a guide to policymaking in the real world. Since 1899 (see *Addyston Pipe & Steel Co. v. United States*) the hostility of federal courts to collusive price fixing and information exchanges which may reduce price competition has been, with only very minor exceptions, consistent and implacable.[8] It seems reasonable to believe that this hostility to collusion is founded on something more than bad economic theory and/or the convenience of judges and prosecutors.

To say the obvious: to the extent that entry is impeded, the presumption in favor of permitting collusion is weakened. When the number of sellers in the market is fixed and immutable, Cournot's treatment of oligopoly, especially in its Vickrey version, has at least a modest claim to plausibility as a guide to policy.[9] (Its claim is no more than modest since Cournot's oligopolists are presumed never to discover any way of exchanging information with one another.)

In the case where entry restrictions are man-made (for example, trucking and FM radio stations), laws against collusion are defensible as a second best policy. (The industry which has the political influence needed to secure entry controls often has enough to gain an exemption from antitrust rules against price fixing but this is another story.) Conceivably the real world is now so riddled with entry barriers created by tax laws, franchise requirements, safety standards, import quotas, zoning regulations, etc. that a presumption in favor of collusion based on a premise of free entry is not strong enough to justify a major change in our present policy.

In this paper my analysis (by Assumption 7) has been restricted to the case where the number of firms is large enough to insure that each will, in equilibrium, have an expected output at which unit cost is minimized.

long-standing cartel and allegedly engaged in a coordinated effort to drive him from the market.

[8]Notably *Appalachian Coals, Inc. v. United States* and *Maple Flooring Ass'n. v. United States*

[9]In the original Cournot model of oligopoly every oligopolist is presumed to know the demand curve for the product of the industry in which he operates. In the Vickrey version every oligopolist begins production in total ignorance of this demand curve. It was Vickrey's achievement to show that each oligopolist, acting alone, would ultimately collect sales data which, when analyzed, would lead him to behave as would a Cournot oligopolist. Let Q_e denote equilibrium output, Q_c the output that would be produced under purely competitive conditions (an infinitely large number of sellers), and S the number of sellers. In both the original and the Vickrey version of the Cournot model of oligopoly, $Q_e = SQ_c/(S+1)$.

Would the welfare conclusions be different if the equilibrium number of firms were "small"? While the issue remains to be investigated, there is no immediately obvious reason why they should be different. Recall the justification offered earlier for restricting our analysis to the case of many firms. We accepted that the firm can have power over price because it sells in a market made imperfect by information costs and/or because it is an oligopolist; and I wished to isolate the effects of these two different types of power. But the generalization must stand that collusion cannot permanently reduce output unless it creates a barrier to entry that would not otherwise exist. In the absence of this result, even for oligopolists, collusion can only increase output in the long run.

Finally, there is the argument that can be made in favor of retaining the standard railroad gauge, nonmetric systems of measurement, the state senate of Rhode Island, or any other anachronism of long standing; the social welfare cost of correcting it may be too high. In some industries the short-run welfare loss that would follow a legalization of collusion would certainly be substantial.

For these reasons it seems best to refrain (at least for the present) from asserting or implying that the analysis of this paper has shown that legalizing collusion will increase economic welfare. Still, there is no reason why my claims for the analysis should be advanced with timidity. The advocates of strong measures against price fixing and information exchanges should no longer be allowed to treat the welfare case for their position as nearly self-evident. It is not. Indeed, they must now come forward and argue a proposition which is, on its face, implausible: that economic welfare can be increased by legal rules that penalize the creation of information about markets. The fact that competition, as measured by the frequency of price change, can sometimes be increased by making producers more ignorant about markets is true but not to the point.

At the very least, my analysis provides ample justification for condemning any use of scarce antitrust enforcement resources to harass small-fry price fixers and low-budget trade associations. Let the local laundries collude in peace.

REFERENCES

P. A. Asch and J. J. Seneca, "Characteristics of Collusive Firms," *J. Ind. Econ.*, Mar. 1975, *23*, 223–37.

——— **and** ———, "Is Collusion Profitable?," *Rev. Econ. Statist.*, Feb. 1976, *68*, 1–11.

Robert H. Bork, *The Antitrust Paradox: A Policy at War with Itself*, New York 1978.

A. A. Cournot, *Recherches sur les principles mathématiques de la théorie des riches*, Paris 1836; trans. by Nathaniel T. Bacon as *Researches into the Mathematical Principles of the Theory of Wealth*, New York 1897.

D. Dewey, "Industrial Concentration and the Rate of Profit: Some Neglected Theory," *J. Law Econ.*, Apr. 1976, *19*, 67–78.

Harvey J. Goldschmid et al., *Industrial Concentration: The New Learning*, Boston 1974.

R. A. Jenner, "An Information Version of Pure Competition," *Econ. J.*, Dec. 1966, *76*, 786–805.

Daniel Marx, *International Shipping Cartels: A Study of Industrial Self-Regulation By Shipping Conferences*, Princeton 1953.

G. W. Nutter and J. H. Moore, "A Theory of Competition," *J. Law Econ.*, Apr. 1977, *19*, 39–65.

Richard A. Posner, *Antitrust Law: An Economic Perspective*, Chicago 1976.

G. B. Richardson, "The Theory of Restrictive Trade Practices," *Oxford Econ. Papers*, Nov. 1965, *17*, 432–49.

George J. Stigler, *The Organization of Industry*, Homewood 1968.

Henri Theil, *Economics and Information Theory*, Chicago 1967.

———, *Applied Economic Forecasting*, Chicago 1966.

William S. Vickrey, *Microstatics*, New York 1967.

Addyston Pipe & Steel Co. v. United States, 175 U.S. 211 (1899).

Appalachian Coals, Inc. v. United States, 288 U.S. 344 (1933).

Maple Flooring Ass'n. v. United States, 268 U.S. 563 (1925).

Paterson Parchment Paper Co. v. Story Parchment Co., 37 F(2d) 537 (1st Cir. 1930).

Differential Inflationary Expectations and the Variability of the Rate of Inflation: Theory and Evidence

By ALEX CUKIERMAN AND PAUL WACHTEL*

Economic models which involve inflationary expectations usually treat them as being uniform across all economic agents in the economy (see, for example, Thomas Sargent and Cukierman, 1977). Obviously, this is dictated more by analytical convenience than by reality. The little directly observed data that exists on inflationary expectations in the United States suggests that even at a point of time people may have widely differing views about the future rate of inflation.[1] Moreover, the distribution of expected inflation across people changes substantially over time.

This paper is an attempt to develop a simple macro-economic framework in which expectations are formed rationally. Nevertheless, people in different markets have different views about the future rate of inflation because they are exposed to different information. This framework is used to explore whether there are some systematic relationships between the variance of expectations across people and the underlying variances of aggregate demand and of the rate of inflation in the general level of prices. The resulting implications are then tested empirically, using data on directly observed price expectations.

The macro-economic model of rational expectations pioneered by Robert Lucas and extended by Robert Barro seems to be a good starting point for an investigation of the variation of expectations. The model incorporates the possibility of divergent expectations in a framework in which market participants know the structure of the economy and form their expectations optimally given this structure and the current price information available to them. Because of lags in the dissemination of aggregate information and an inability to distinguish immediately between aggregate and specific causes for movements in the price of their own good, suppliers in different markets will have different views about the general level of prices. Lucas and Barro attribute differences in information (and consequently expectations) to the existence of physically separate markets for a single aggregate good. However, as discussed below, their models will apply also when markets for different goods are physically near to each other, but their participants possess different information about prices in the economy. This could arise because a regular participant in a given market has a cost advantage in obtaining up-to-date information on the price in that market over obtaining information on prices in other markets in which he is not usually engaged. We exploit this way of looking at markets here. The formal model developed is a variant of Lucas' earlier work.

Section I presents a modified version of the Lucas model. In Section II the implications of rationality and information differences across markets for the spread of inflationary expectations are worked out. In particular, it is shown that an increase in the variance of the rate of change of nominal income is likely to increase both the variance of inflationary expectations across markets and the variance of the actual rate of inflation. The effect of a change in the variance of relative demand shocks on the variance of inflationary expec-

*Carnegie-Mellon University and Tel-Aviv University, and New York University Graduate School of Business Administration, respectively. Our major debt is to Robert Lucas. We benefited from comments made by participants at various workshops, as well as suggestions made by Edi Karni, the referees, and the managing editor of this *Review*. Thanks are due to Biplab Das for research assistance. Any remaining errors are, of course, our responsibility.

[1]There are two sources of these data—Livingston's survey of forecasters and the University of Michigan Survey Research Center (*SRC*) surveys. The distribution of expectations from these surveys is discussed by Wachtel and John Carlson (1977).

tations is investigated as well. Section III presents some empirical tests of the implications derived in Section II, using data on the variance of expected inflation across people computed from the Livingston and *SRC* data.

I. The Model

The model differs from that presented by Lucas in several elements. First, it is interpreted as a many-goods model, rather than a single good model with trading at separate locations. Second, and most importantly, there are as many equilibrium conditions as there are goods or markets. An important implication of this modification is that in contrast to the original Lucas model, the variance of relative prices is not constant, but is systematically related to the variance of the general price level, as well as to some exogenous variances to be introduced shortly.[2] Third, a distinction is drawn between a market and the realization of market-specific shocks in that market. As a consequence, the general level of prices is naturally defined as a weighted average of prices in individual markets. Fourth, we abstract from the effect of lagged or actual employment which appears in Lucas. A short description of the model and its equilibrium characteristics follows.

Trading in the different goods is scattered among a large number of competitive markets. Each of these markets may be taken as the market for different goods or the same good at different locations. An essential feature of this market separation is that the flow of information among markets is not instantaneous. Although suppliers have full information about the current price in their own market, their information about prices in other markets and consequently, about the general price level, is incomplete.[3]

A. *Supply*

Quantity supplied in each market is the product of a normal or secular component that is common to all markets and a cyclical component that varies from market to market. Letting v index markets and using $y_{nt}(v)$ and $y_{ct}(v)$ to denote the *logs* of these components, supply in market v is

$$(1) \qquad y_t(v) = y_{nt} + y_{ct}(v)$$

The secular component follows the trend line

$$(2) \qquad y_{nt} = \alpha + \beta t$$

The cyclical component varies with perceived relative prices

$$(3) \quad y_{ct}(v) = \gamma[\, p_t(v) - E(Q_t \,|\, I_t(v)] \quad \gamma > 0$$

where $p_t(v)$ is the *log* of the actual price in market v at time t and $E(Q_t \,|\, I_t(v))$ is what suppliers in market v believe the mean current (*log* of the) general price level to be, based on the information available to them in this market. The information $I_t(v)$ available to individuals in market v at time t includes both information which is common to all markets[4] and market-specific information summarized in the equilibrium price observed in that market at time t. Equation (3) states that supply in market v is an increasing function of

[2]This modification which is due to Lucas was suggested by him in private correspondence. It is preferred for two reasons. First, it implies a positive relationship between the variance of relative prices and some attributes of aggregate variability (see Cukierman, 1979), a relationship which has been found to hold empirically by Daniel Vining and Thomas Elwertowski. Second, only with it does the central implication of Lucas' paper follow unambiguously, that the short-run Phillips tradeoff deteriorates when the variance of the rate of change in nominal income increases.

[3]Information gaps can occur even if the markets are physically close. For example, wheat dealers and car dealers in the same city will find the cost of gathering information to be less in the market in which they participate regularly. Although some general price information will be available to everyone, it is likely to be less timely than the information about one's own market. For example, newspaper publication of the price indices conveys information about other markets which is available at very little cost. However, given the substantial delays between data collection and publication, that information becomes available with less immediacy than the information about the price in one's own market.

[4]This common information is summarized by the expected value of the general price level in the current period based only on information about the economy prior to the current period ($\overline{Q}_t$). This information is disseminated among markets in various ways, including a look at newspaper reports on price indexes.

the relative price of good v as perceived by the suppliers of this market.

B. *Demand*

Demand in each market is stochastic. The stochastic component has two parts: one common to all markets and the other specific to the particular market. The demand function is

$$y_t(v) + p_t(v) = x_t + w(v) \tag{4}$$

where x_t is an exogenous random shift variable that is common to all markets and $w(v)$ is a random shock whose realization is specific to market v but whose distribution is common to all markets. By definition $x_t = x_{t-1} + \Delta x_t$. In period t, x_{t-1} is already determined and known by the participants in all markets. However, Δx_t is random. It is assumed that the random shocks are distributed normally and are independent of each other.

$$\Delta x_t \sim N(\delta, \sigma_x^2) \tag{5a}$$

$$w \sim N(0, \sigma_w^2) \tag{5b}$$

The sum of demands (from (4)) over all markets is equal to total nominal income. It is shown in Appendix B that when the number of markets is large, Δx reflects the rate of change in the level of aggregate nominal income and therefore σ_x^2 can be interpreted as the variance of the rate of change in nominal income.

C. *The General Price Level and its Distribution*

The general price level Q_t is defined as a fixed-weight index of the prices in individual markets, with weight $u(v)$ assigned to market v.

$$Q_t \equiv \sum_v u(v)p_t(v) \tag{6}$$

$$\sum_v u(v) = 1 \qquad u(v) \geq 0 \text{ for all } v$$

It is further assumed that for any v the weight $u(v)$ assigned to market v is small—(see Appendix A) in comparison to the sum of all the weights. This assumption is crucial in assuring later that the general price level would depend only on the realization of the economy-wide shock and not on the realization of the market-specific shocks.[5] Note that even though the effect of market-specific shocks on Q_t will turn out to vanish, Q_t still remains an unknown stochastic variable in market v at period t. This is because in period t the realization of the random common demand shock x_t is still unknown in market v, and this shock does affect the general level of prices.

D. *Rational Expectations and Equilibrium*

Expectations regarding the general price level are formed rationally in the sense that given the currently available information, participants in each market use the structure of the economy, which is known to everyone, to form optimal forecasts. Furthermore, actions based on these forecasts do in fact generate the assumed economic structure.

In our model, the beliefs of market participants can be represented by equations (7)–(10):

$$p_t(v) = Q_t + z(w) \tag{7}$$

where

$$Q_t \sim N(\bar{Q}, \sigma^2) \tag{8}$$

$$z(w) \sim N(0, \tau^2) \tag{9}$$

and where $z(w)$ is the percentage deviation of the price in a particular market from the general price level, which depends only on the realization of the stochastic term w. Furthermore, the expected price level from the standpoint of information available in market v is given by

[5]By contrast, Lucas (p. 328) defines the general price level implicitly as the expected value of price in any given market across the distribution of market-specific demand shocks. The definition used here is more attractive for two additional reasons First, it allows the allocation of different weights to different markets in the computation of the general price index. This is particularly important if the model is to be interpreted as a many-goods model in which the weights of the different goods are not necessarily equal. Second, since these weights can be thought of as relative market shares for the case in which the realization w is equal to zero (its expected value), the price index is essentially similar to the standard price indices.

$$(10) \qquad E[Q_t | I_t(v)] = (1-\theta)p_t(v) + \theta \overline{Q}_t$$

where $$\theta = \frac{\sigma_w^2}{\sigma_x^2 + \sigma_w^2}$$

We will demonstrate below that the actual behavior of the economy will also be described by these equations. At this stage we shall assume that (7)–(10) are satisfied and proceed to solve for the equilibrium aggregate price level on the further assumption that each market is cleared. It will be shown that (7)–(10) are in fact satisfied in equilibrium.[6] Substituting (10) into (3), (2) into (1), and equating the resulting quantity supplied in market v with the quantity demanded in that market (from (4)), we may solve for the price in market v as

$$(11) \quad p_t(v) = \frac{1}{1+\theta\gamma} \cdot [x_t + w(v) - \alpha - \beta t + \theta\gamma\overline{Q}_t]$$

As can be seen from (11), equilibrium price on market v depends on the realization of the relative demand shock w in that market but does not depend on the market index v. This is a consequence of the uniform demand and supply structure assumed for all markets. As a result, prices in different markets will differ only to the extent that the realizations of the relative demand shocks w differ among markets. We shall therefore use the notation $p_t(w)$ to denote equilibrium price in a particular market.

Substituting (11) into (6) the realization of the general price level may be written

$$(12) \quad Q_t = [x_{t-1} + \delta + \gamma\theta\overline{Q}_t - \alpha - \beta t + \epsilon_t + \sum_v u(v)w(v)] \div (1+\gamma\theta)$$

where use has been made of the identities $x_t = x_{t-1} + \Delta x_t = x_{t-1} + \delta + \epsilon_t$, and ϵ_t is defined as the stochastic deviation of Δx_t from its mathematical expectation δ, and is unknown in period t. It has the same normal distribution as Δx_t except that it is distributed around 0 rather than around δ. If the weights $u(v)$ which enter into the general price level index in (6) are all small in comparison to their sum, Chebyshev's inequality implies that the probability that the summation term in the numerator of (12) is different from zero tends to zero as the number of goods (markets) tends to infinity.[7] Assuming that the number of goods is large we may therefore neglect the summation term in the numerator of (12). By taking the expected value over ϵ_t of (12) we obtain

$$(13) \qquad \overline{Q}_t = x_{t-1} + \delta - \alpha - \beta t$$

Substituting (13) into (12) we note that

$$(14) \qquad Q_t = \overline{Q}_t + \frac{\epsilon_t}{1+\theta\gamma}$$

which implies that

$$(15) \qquad Q_t \sim N(\overline{Q}_t, \sigma^2)$$

where $$\sigma^2 = \frac{\sigma_x^2}{(1+\theta\gamma)^2}$$

This establishes (8). It is worth noting that producers in all markets are assumed to know all the parameters of the model, as well as the value of last period's cumulated common demand shock, x_{t-1}. They can therefore compute $\overline{Q}_t$ from (13). However, Q_t will in general be different than $\overline{Q}_t$ because the aggregate demand shock in period t may deviate from its expected value and this becomes known only in the next period.

E. *Equilibrium Prices on Individual Markets*

The equilibrium price on any particular market can be rewritten by substituting (12) into (11):

$$(16) \quad p_t(w) = \frac{1}{1+\theta\gamma}[x_t - \alpha - \beta t + \theta\gamma\overline{Q}_t] + \frac{w}{1+\theta\gamma} = \overline{Q}_t + z(w)$$

That is, the price on any particular market is equal to the general price level plus a

[6]Since the model involves fewer terms than the Lucas model, it is easier to solve it directly, which we do. By contrast, Lucas and Barro use the method of undetermined coefficients. Obviously, the result is the same whatever the solution method.

[7]For details and a proof, see Appendix A.

deviation z, *which depends only on* w *and whose explicit form is* $w/(1+\theta\gamma)$. *This* establishes the validity of (7). Furthermore, (16) and (5b) imply that

$$z(w) \sim N(0, \tau^2) \tag{17}$$

where

$$\tau^2 = \frac{\sigma_w^2}{(1+\theta\gamma)^2}$$

which proves (9). Note also that the relative price is given by $p_t(w) - Q_t$, so that τ^2 is also the variance of relative prices.

F. *The Formation of Expectations*

In period t producers in each market know the past history of the general price level. From their knowledge of Q_{t-1} they can compute x_{t-1} (using (12) lagged one period) with which they can then compute $\overline{Q}_t$ using (13). Producers in each market also know the current price $p_t(w)$ in their market and since they know the structure of the model, they know that it partially reflects (through (16)) movements in Q_t. By (15), (16), and (17), $p_t(w)$ is also normally distributed with known expected value and variance. So a producer in any given market is confronted with the need to form the best possible prediction of the currently prevailing general price level Q_t, conditional on his knowledge of $\overline{Q}_t$ and $p_t(w)$. If he forms his prediction optimally by minimizing the expected mean square error of the prediction, given $p_t(w)$ and $\overline{Q}_t$, statistical theory implies that the optimal forecast is[8]

$$(18)\quad Q_t^*(w) = E[Q_t \mid p_t(w), \overline{Q}_t] = \frac{\sigma_x^2}{\sigma_x^2+\sigma_w^2}p_t(w) + \frac{\sigma_w^2}{\sigma_x^2+\sigma_w^2}\overline{Q}_t$$

$$Q_t^*(w) = (1-\theta)p_t(w) + \theta\overline{Q}_t$$

This proves (10) and hence establishes the consistency of the model. More precisely, we solved for the equilibrium of the model by assuming that (7)–(10) are fulfilled and showed that this equilibrium, together with the hypothesis that people form their expectations optimally, imply the existence of (7)–(10). That is, the structure of prices and of expectations which initially were assumed to be given were later obtained as conclusions using the assumptions of equilibrium in the market for each good and of optimality in the formation of expectations.[9]

It is interesting to note that this model is perfectly consistent with an information structure in which individuals in different markets all get information on past values of the general price level from the periodic publication of price indices (see fn. 3). In terms of this model, participants in the various markets all know the prior value of the general price level from newspaper reports. They then use this information in the way outlined above in order to infer $\overline{Q}_t$ which in turn is used as one of the inputs in forecasting Q_t^* by means of (18). However since observation of $\overline{Q}_t$ does not contain information about the current aggregate shock ϵ_t, while the price of an individual good does, a producer in any given market may sharpen his forecast by using the information on the price of his own product as well as the past price information that he gets from the mass media.[10]

II. The Distribution of Expectations Across Markets

An interesting implication of the above model and one which is central to this paper is that although expectations are formed rationally, they are not necessarily identical across markets. The reason for different expectations is that producers in the various markets are confronted with different bits of information. Since the equilibrium price level on any

[8]For the normal distribution, the point estimate that minimizes the mean square error of prediction, given $\overline{Q}_t$ and $p_t(w)$, is equal to

$$\rho_{Q_t p_t(w)} \frac{[p_t(w) - Ep_t(w)]}{\sigma_{p_t(w)}}\sigma_{Q_t} + EQ_t = \frac{\sigma_x^2(p_t(w)-\overline{Q}_t)}{(\sigma_x^2+\sigma_w^2)} + \overline{Q}_t$$

where ρ denotes a coefficient of correlation. See H. D. Brunk (pp. 212–18), for example. Application of that theorem to our case yields (18).

[9]For an intuitive discussion of the determinants of the weights θ and $(1-\theta)$ in (18), see Lucas (p. 328).

[10]There may also be lags in the assimilation of economy-wide information on past general price indices. In any case, the combined effect of the lags in the dissemination of reports on the general price level cannot be expected to be very long.

given market, $p_t(w)$, varies with the actual realization of w in that market, producers in different markets are generally confronted with different prices. Since they all have the same information about the structure of the economy and about the expected value of the price level, they all pool the information on their own price and on $\overline{Q}_t$ in an identical manner as suggested by equation (18). However, this results in different point estimates of the expected aggregate price level (Q_t^*), even though the perceived distributions are the same in all markets. Equation (18) also suggests that the expected general price level is different across markets only when the realizations of the relative demand shocks are different across markets. In addition, (18) suggests that for a given distribution of the w's across markets, the spread of expectations in all markets will be larger, the larger $(1 - \theta)$ or the larger is the variance of the aggregate demand shocks (σ_x^2) in comparison to the variance of relative demand shocks (σ_w^2). Appendix B shows that as long as the number of markets is large, σ_x^2 can also be viewed as the variance of the rate of change in aggregate nominal income. This leads to the following proposition:

PROPOSITION 1: *A ceteris paribus increase in the variance of the rate of change of nominal income increases the variance of the expected general price level across different markets as well.*

This proposition suggests that the higher the amount of "noise" in the system, whether caused by nature or by the willful policy of the monetary authority, the higher will be the divergence of views about the current price level.

A. *The Distribution of the Expected Rate of Inflation Across Markets*

We shall measure the expected rate of inflation in period t as the percentage difference between the general price level forecast for period (t+1) and the general price level forecast for period t where both forecasts are made in period t in a particular market. That is,[11]

$$(19) \quad \pi_t^*(w) = E[Q_{t+1} \mid I_t(w)] - E[Q_t \mid I_t(w)]$$

Equation (14) with a lead of one period implies

$$(14') \qquad Q_{t+1} = \overline{Q}_{t+1} + \frac{\epsilon_{t+1}}{1 + \theta\gamma}$$

Substituting for $\overline{Q}_{t+1}$ from (13) with a lead of one period and taking expectations yields

$$(20) \quad E[Q_{t+1} \mid I_t(w)] = \overline{Q}_t + \delta - \beta + E[\epsilon_t \mid I_t(w)]$$

The information available in a particular market at time t includes $p_t(w)$ which partially reflects movements in ϵ_t (see (16)). Thus, the last term in (20) is nonzero, and the formula in footnote 8 is used to compute the optimal prediction of ϵ_t given $p_t(w)$ which yields

$$(20') \quad E[Q_{t+1} \mid I_t(w)] = \overline{Q}_t + \delta - \beta + (1 - \theta)(\epsilon_t + w)$$

Substituting (18) and (20') into (19), and using (14) and (16), we obtain

$$(21) \quad \pi_t^*(w) = \delta - \beta + \frac{\gamma\theta(1 - \theta)}{1 + \gamma\theta}(\epsilon_t + w)$$

as the expression for the rate of inflation expected in the general level of prices by people in a market with a relative demand shock w at time t.[12] The mathematical expectation of the expected rate of inflation across the distribution of w is given by

$$(22) \quad \underset{w}{E}\pi_t^*(w) = (\delta - \beta) + \frac{\gamma\theta(1 - \theta)\epsilon_t}{1 + \gamma\theta}$$

The variance of the expected rate of inflation across the distribution of w is[13]

[11]Note again that it depends on w, rather than on market index v, since informational differences among markets arise only because of differences in w.

[12]Equation (21) implies that inflationary expectations are distributed normally. Carlson (1975) tests this hypothesis using the Livingston data and is not able to reject the hypothesis of normality.

[13]Note that the variance of the expected rate of inflation is time independent as long as σ_x^2 and σ_w^2 are time independent. It is also interesting to note that the variance in (23) is not only the variance of one-period-ahead inflationary expectations, but the variance of the jth ($j = 1, 2, 3, \ldots$) period-ahead inflationary expecta-

(23) $$V(\pi^*) = \underset{w}{E}[\pi_t^*(w) - \underset{w}{E}\pi_t^*(w)]^2 = \gamma^2\sigma_w^2\left[\frac{\theta(1-\theta)}{1+\gamma\theta}\right]^2$$

PROPOSITION 2: *The variance of the expected rate of inflation increases when the variance of the rate of change in nominal income increases if and only if* $r = \sigma^2/\tau^2 < (1+\gamma)^{1/2}$, *where r is the ratio of the variance of the general price level to the variance of relative prices, and* γ *is the elasticity of supply with respect to the perceived relative price in any given market.* (See Appendix C for the proof.)

The elasticity of supply is nonnegative so the minimum value of $(1+\gamma)^{1/2}$ is one. Thus, the positive relationship between the variance of the expected inflation rate and the variance of the rate of change in nominal income holds whenever $r = \sigma^2/\tau^2 < 1$. However, depending on the size of the supply elasticity γ, it will also hold when σ^2, the variance of the general price level, is greater than the variance of relative prices, τ^2. For example, when $\gamma = 1$, Proposition 2 is valid as long as the variance of the price level is less than about 1.4 times the variance of relative prices. For $\gamma = 10$, the critical value, $(1+\gamma)^{1/2}$ reaches 3.16. In the empirical section we present evidence which suggests that the sufficient condition for Proposition 2 to hold is likely to have been satisfied in the postwar period.

B. *The Distribution of the Actual Rate of Inflation*

The actual rate of inflation in the general price level is by definition,

(24) $$\pi_t = Q_t - Q_{t-1}$$

The mathematical expected value of π_t over the distribution of ϵ is equal to $\overline{Q}_t - \overline{Q}_{t-1}$ from (14). The corresponding variance is given by

(25) $$V(\pi) = \underset{\epsilon}{E}[(Q_t - \overline{Q}) - (Q_{t-1} - \overline{Q}_{t-1})]^2 = 2\sigma^2 = 2\frac{\sigma_x^2}{(1+\theta\gamma)^2}$$

C. *The Relationship between the Variance of the Expected Rate of Inflation and the Variance of the Rate of Inflation*

A comparison of $V(\pi)$ and $V(\pi^*)$ suggests that they have many elements in common which raises the possibility that when one variance changes, the other does as well, perhaps even in the same direction. The following proposition establishes the exact relationship between the variance of the rate of inflation and the variance of the expected rate of inflation across market participants. Proposition 3a deals with the case of an exogenous change in σ_x^2 and Proposition 3b deals with the case of a change in σ_w^2.

PROPOSITION 3: a) *A ceteris paribus change in the variance of the rate of change in nominal income* σ_x^2 *changes the variance of the rate of inflation and the variance of expected inflation in the same direction if and only if* $r < (1+\gamma)^{1/2}$. b) *A ceteris paribus change in the variance of relative demand shocks,* σ_w^2, *changes the variance of the rate of inflation and the variance of expected inflation in the same direction if and only if the following conditions on* $r = \sigma^2/\tau^2$ *hold:*

$$r < \frac{3\gamma + 4 - (\gamma^2 + 16(1+\gamma))^{1/2}}{\gamma - 4 + (\gamma^2 + 16(1+\gamma))^{1/2}} = r_D(\gamma)$$

$$\textit{or} \quad r > \frac{3\gamma+4}{\gamma-4} \quad \textit{and} \quad \gamma > 4$$

(See Appendix D for the proof.)

Proposition 3a implies that if σ_x^2 changes with σ_w^2 constant and $r < (1+\gamma)^{1/2}$, then the variance of the rate of inflation will be positively related to the variance of the expected rate of inflation over different markets. Proposition 3b implies that if σ_w^2 changes as well, this positive relationship will be enhanced when $r < r_D(\gamma)$ and weakened or possibly

tions as well. Therefore, Propositions 2 and 3 in the text hold for any forecast horizon. This is a consequence of the fact that our model does not allow for a change in uncertainty when the forecast horizon is extended.

reversed when $r > r_D(\gamma)$.[14] The critical ratio r_D is an increasing function of γ provided the condition of Proposition 3b is satisfied.[15] Furthermore, $r_D(\gamma)$ lies between zero and one for any positive γ. Since $(1 + \gamma)^{1/2}$ is always larger than one, it follows that the condition in Propositions 2 and 3a is less stringent than the condition in Proposition 3b. In other words, if the condition for a positive relationship between $V(\pi)$ and $V(\pi^*)$ when σ_w^2 changes is valid, then so is the condition which assures a positive relationship between those two variances when σ_x^2 changes.

The results of the last two propositions can be understood intuitively as follows: An increase in the variance of the rate of inflation which is caused by an increase in σ_x^2 comes about concomitantly with an increase in the variance of relative prices.[16] The increase in the variance of relative prices makes the spread of $p_t(w)$ across individual markets greater, which tends to increase the spread of expectations concerning both the current and the future price levels (through (18) and (20)). This effect is reinforced by the fact that the weight which is given to market-specific information increases when σ_x^2 increases. Provided $r < (1 + \gamma)^{1/2}$ the increase in the spread of expectations concerning both the current and the future price level also causes an increase in the spread of inflation expectations.[17]

[14]This statement is based on the presumption that $\gamma \leq 4$.

[15]This condition is needed to assure this result only for $\gamma > 4$ and even then it is only a sufficient and not a necessary condition. In the empirical section the evidence on whether $r < r_D$ will be examined.

[16]The variance of relative prices unambiguously increases when σ_x^2 increases. This is shown in Cukierman (1979, equation (5)).

[17]The condition $r < (1 + \gamma)^{1/2}$ is needed to assure that the decrease in the (negative) covariance between the expectations regarding the current and the future price levels caused by the increase in σ_x^2 is smaller in absolute value than the increase in the variances of those two expectations across markets It is interesting to note in this context that there is empirical evidence which suggests a positive relationship between the variance of relative price change and the variance of the rate of inflation (see Vining and Elwertowski) as well as empirical evidence which suggests a positive relationship between the variance of inflationary expectations and the variance of relative price change (see our unpublished paper).

III. Empirical Tests

The empirical tests focus on the relationships contained in Propositions 2 and 3 between the variance of nominal income and price change and the variance of inflationary expectations across markets. Empirical proxies for the variances of the inflation rate and the rate of change of nominal income will be defined below. Data on the variance of the expected rate of inflation are obtained from two surveys—Livingston's survey of forecasters and the *SRC* consumer survey.

Joseph A. Livingston is a newspaper columnist who has conducted a semiannual survey of forecasters and business economists since 1947. Usually the number of survey respondents is between thirty and seventy-five. A consistent time-series for the expected rate of inflation in the Consumer Price Index (*CPI*) has been constructed from the original survey responses. In this paper we use the variance across survey respondents of the twelve-month expected inflation rate in the *CPI*.[18]

The *SRC* at the University of Michigan includes in its *Quarterly Consumer Survey* a question about inflationary expectations. Until 1966 the survey only obtained qualitative information about the direction of expected price change (up, down, or remain the same). Starting in 1966, respondents expecting price increases were also asked how much. In recent years the question has been improved to allow respondents to provide an open-ended response about the amount of expected inflation. The question refers to consumer prices generally and inflation over a one-year span. The number of respondents varies from survey to survey but it is always at least several hundred and is often more than one thousand. Finally, unlike the Livingston survey, the respondents are representative of the population at large.

The post-1966 data which include quantitative responses can be used to calculate a time-series for the variance of the expected

[18]The data are Carlson's calculations (1977, Table AI, p. 53). The movements over time in the survey variances of six-month inflation forecasts of the *CPI* and of forecasts of the Wholesale Price Index are broadly similar and therefore not shown here.

inflation rate across survey respondents. The series used here was constructed by F. Thomas Juster and Robert Comment. It uses responses from the open-ended question introduced in 1973 and corrects the earlier responses for bias.[19]

These data on the variance of the expected rate of inflation across survey respondents at a point of time are a natural counterpart for the theoretical variance $V(\pi^*)$ (which it will be recalled is the variance of the expected rate of inflation across the theoretical distribution w). The identification of $V(\pi^*)$ with the survey variance is based on the assumption that the surveys randomly sample people in different markets or information centers.[20] At a moment of time $V(\pi^*)$ is given and consequently the sample observations of the expected rate of inflation, at this time, can be taken as drawings from a distribution of expected rates of inflation with theoretical variance $V(\pi^*)$. The sample variance will then be an estimate of this theoretical variance.

The empirical proxies for the theoretical variances of the actual rate of inflation and the rate of change in nominal income are their variances over time. For the theoretical variance of the actual inflation rate, variances of the rate of change in the *CPI* (monthly data) and in the *GNP* deflator (quarterly data) were constructed. For the actual rate of change in nominal income, proxies were constructed from the rates of change of *GNP* (quarterly) and personal income (monthly). The variance of the rate of change of *GNP* is a good proxy for the theoretical variance, since when averaged over all the goods in *GNP*, the relative disturbances w average out to zero so that the rate of change in total *GNP* equals the rate of change in the common component of nominal income (Δx) in all markets.

Before we turn to the empirical investigation of the relationship of $V(\pi^*)$ with $V(\pi)$ and σ_x^2, recall that when these relationships were derived in Propositions 2 and 3, they were conditioned upon relationships between σ^2 and τ^2. We now examine the available evidence which does indicate that the conditions sufficient for the propositions to hold are fulfilled. Propositions 2 and 3a suggest that if the changes over time in $V(\pi^*)$ and $V(\pi)$ are caused by changes in σ_x^2, a sufficient condition for a positive relationship between $V(\pi^*)$ on one hand, and either of $V(\pi)$ or σ_x^2 on the other, is that $\sigma^2 < \tau^2$. That is, if σ^2, the variance of the general price level, is less than τ^2, the variance of relative prices, Propositions 2 and 3a of our model are valid.

In order to determine whether this condition is fulfilled, we compared several alternative measures of the moving variance of the inflation rate to the variance of relative price change computed by Vining and Elwertowski from annual data for the components of the *CPI*, and by Richard Parks from the annual data for major components of the implicit price deflator (*IPD*) for consumption expenditures.[21] These are compared to five-year centered moving variances of the annual rates of change of the *CPI* and the implicit price deflator, respectively. For the Vining-Elwertowski *CPI* comparison, $r = \sigma^2/\tau^2$ is always less than unity and, except for the first four years of the sample period (1948–74), it is less than .15 in all but one year. For the Parks *IPD* comparison, it is less than unity in all but six years of the sample period (1947–75), four of which are prior to 1951.[22] The Vining-

[19]Although Juster and Comment provide estimates of the survey variance for the period 1948–66, inferred from the distribution of the qualitative responses, we do not use them here. These data are clearly a less reliable measure and the variance estimates are quite sensitive to changes in the procedures used.

[20]This assumption is more appropriate for the *SRC* data than the Livingston data since only the former is based on a random sample. However, it is only necessary to assume that the variability of expectations among the forecasters surveyed by Livingston is positively correlated with the true variability for our test to be valid with the Livingston variance as well.

[21]These studies provide observations on the variance of relative price change which in our terms is equal to $\tau_{t+1}^2 + \tau_t^2$. However, our measures of the moving variance of the rate of inflation are estimates of the variance of the rate of inflation, $V(\pi_t) = \sigma_{t+1}^2 + \sigma_t^2$, rather than an estimator of the variance of the price level (σ_t^2). The ratio between this moving variance and the Vining-Elwertowski variance approximates the ratio $r = \sigma^2/\tau^2$. This can be seen by noting that in terms of the model, we compute $(\sigma_{t+1}^2 + \sigma_t^2)/(\tau_{t+1}^2 + \tau_t^2)$ which amounts to the estimation of the average of r over two consecutive periods.

[22]The results are less satisfying with the Parks data because the Parks measure of τ^2 understates the variability of relative price change. It is based on only twelve

Elwertowski measure of τ^2 is also larger than (with two exceptions early in the period) the variance during each year of the monthly rates of change in the *CPI* at annual rates. This is a more stringent test of the condition that $r = \sigma^2/\tau^2 < 1$ because there is more variability in monthly than yearly inflation rates. Since, in addition, the condition $r < 1$ is sufficient and necessary only for the extreme case $\gamma = 0$ and becomes less stringent as γ increases, we conclude that the empirical evidence supports the view that the conditions sufficient for the validity of Propositions 2 and 3a are satisfied. We should therefore expect to find a positive relationship between the survey variances on one hand and the moving variances of the rate of change in nominal income and of the rate of inflation on the other.

Comparisons of the survey variance with the variances of nominal income and price change are made in two ways. The first compares the average survey variance of inflationary expectations in a given time span to the variance in the actual inflation rate and in the rate of change of nominal income in that span. The second method examines the correlation between the survey variance and a moving variance of the actual inflation rate or the rate of change of nominal income. Both are used to test the positive relationship between the variance of the rate of inflation or the variance of the rate of change in nominal income on one hand and the survey variance of the expected rate of inflation on the other as suggested by Propositions 2 and 3.

The Livingston data are used to apply the first method since it is available for a longer period. The postwar period was broken down into several subperiods, each with a substantially different variance of the actual rate of inflation. The average survey variance of expected inflation was computed for the same time periods and compared with the variances of the rate of inflation and of the rate of nominal income change for those periods. Table 1 shows the comparison of the average variance from the Livingston surveys and the variances over time in the quarterly rates of change of *GNP* and of the *GNP* deflator.

TABLE 1—AVERAGE VARIANCES OF LIVINGSTON EXPECTED INFLATION RATE, NOMINAL INCOME CHANGE, AND ACTUAL INFLATION RATE

Period	Variance of Expected Rate of Inflation	Variance of Rate of Change in Nominal *GNP*	Variance of Rate of Inflation in *GNP* Deflator
6/47–6/53	41.36	69.44	22.12
12/53–12/58	3.35	27.70	2.60
6/59–12/65	0.74	15.57	1 31
6/66–6/71	2.30	8 88	1.68
12/71–12/72	1.18	8.73	2.38
6/73–12/75	7.36	20.61	8.63

Note: The first column is the average variance of the Livingston expected inflation rate from surveys conducted in the time period indicated; surveys are conducted in June and December. The other columns are the variances of quarterly percentage rates of change at annual rates beginning with the quarter after the first survey (for example, the first subperiod includes the quarters 1947III through 1953IV).

The results suggest that there is a positive relationship between the variance of the expected rate of inflation across people and both the variance of the rate of change of nominal *GNP* and the variance of the rate of inflation in the *GNP* deflator. There seems to be a clearer positive relation between these variances at the extremities (i.e., for high and low variances) than for intermediate values of the variances. Although these results do not give a very good picture of the strength of the association between the variances, they do point out that a positive association exists.[23]

Since the above results may well depend on the particular breakdown into subperiods chosen and provide no test of the strength of the relationship, a method that correlates the two types of variances is called upon. This is achieved by computing moving variances of the rate of inflation and the rate of change in nominal income. At each point in time the moving variance is a variance in the appropriate rate of change over a two-year span

major components of the price index, while the Vining-Elwertowski variance measures the dispersion among several hundred items in the *CPI*.

[23]Similar results are obtained with the variances of personal income and consumer price change.

TABLE 2—REGRESSIONS OF VARIANCE OF INFLATIONARY EXPECTATIONS ON MOVING VARIANCE OF NOMINAL INCOME CHANGE AND INFLATION RATE

	Time Period	Constant	Slope	R^2
Inflationary Expectations from Livingston Data:				
Consumer Price Index	6/48–12/75	.329	.181	.593
		(.8)	(8.9)	
GNP Deflator	6/48–12/75	.644	.312	.519
		(1.5)	(7.6)	
GNP	6/47–12/75	1.558	.065	.086
		(1.5)	(2.3)	
Personal Income	6/47–12/75	−.089	.041	.824
		(−.2)	(16.2)	
Inflationary Expectations from *SRC* Data:				
Consumer Price Index	1966II–76I	27.736	2.266	.407
		(6.5)	(5.1)	
GNP Deflator	1966II–76I	24.905	6.102	.475
		(5.9)	(5.9)	
GNP	1966II–76I	17.121	2.07	.400
		(2.8)	(5.0)	
Personal Income	1966II–76I	46.759	−.080	.006
		(6.4)	(−.5)	

Note: *t*-statistics are in parentheses below regression coefficients. All variables are measured as percentage changes at annual rates.

centered on the current period.[24] These are then compared to the survey variance of the expected rate of inflation for the corresponding period. Only those observations for which survey data are available are included in the comparison.

The procedure provides an observation, based on the moving variance, of the variance of actual inflation or the rate of change in nominal income, for each survey observation. It is therefore possible to evaluate the strength of the relationship between those variances by running regressions of the survey variance on the various moving variances. Results using both the Livingston and *SRC* variances of the expected rate of inflation are shown in Table 2.

It is clear from Table 2 that there is a significant positive association between the variance of actual inflation or the variance of nominal income and the variance of the expected rate of inflation. With the Livingston variance of inflationary expectations the slope coefficient is always significant. The *SRC* variance, which is only available for a shorter period, yields a similar picture when correlated with the variances of the rate of change in the two price indices and *GNP*, but is not significant when regressed on the variance of the rate of change in personal income.

The slope coefficients vary substantially among the regressions in Table 2. In particular, the slope coefficients of the regressions which use the *SRC* data are typically larger than the coefficients of the regressions which use the Livingston data. This is mostly a reflection of the larger size of the *SRC* data variances. The *SRC* variances are larger for two reasons: First, those variances are measured over a more informationally heterogeneous group than the Livingston variances. Secondly, they refer only to the period since 1966. There is also a systematic difference in the size of the coefficients between regressions which use the variance of the rate of change in nominal income and those which

[24]For the data available monthly (*CPI* and personal income), the moving variance is based on twenty-five observations of the rate of change, while for the quarterly data (*GNP* and *GNP* deflator), only nine are included. In each case, a centered two-year span is used and the monthly or quarterly percentage rates of change are on an annual basis without compounding. A moving variance of the inflation rate has been used by Benjamin Klein to estimate the effect of price uncertainty on demand.

use the variance of the rate of inflation. For both measures of the variance of inflationary expectations, the slope coefficients are always smaller when the variance of the rate of change in nominal income is used in a regression. This is caused by the fact that the variance of the rate of change in nominal income is equal to the sum of the variance of the rate of inflation, the variance of the rate of change in real income and their covariance. Since the covariance, a Phillips curve type of relationship, is likely to be positive, it is not surprising that the variance of the rate of change in nominal income exceeds the variance of the inflation rate, making the coefficient on the latter variance larger.

The large differences in explanatory power among the regressions with the Livingston and *SRC* variances and the variances of the rate of change in nominal income are due to differences in the time periods. In fact, when the Livingston regressions are estimated with data from the last decade of the sample, the results (not shown in Table 2) are very much similar to those shown with the *SRC* data. That is, the R^2 for the *GNP* equation is more than twice as large as that in the third row of Table 2 and the R^2 for personal income (fourth row) drops dramatically towards zero. Nevertheless, the results do lend support to the implications developed in Propositions 2 and 3.

In general, the regressions with the variance of the rate of inflation yield stronger relationships with $V(\pi^*)$ than those with the variance of the rate of change of nominal income. A possible explanation is that if the condition of Proposition 3b is fulfilled and if σ_w^2, as well as σ_x^2, changes over time, then the relationship between $V(\pi^*)$ and $V(\pi)$ is strengthened in comparison to a situation where σ_w^2 is held constant. Since this element is absent from the relationship between $V(\pi^*)$ and σ_x^2, it is not surprising that those regressions tend to be weaker. The evidence discussed earlier on r suggests that the stringent condition of Proposition 3b is valid. Even for a relatively small supply elasticity, $r_D(\gamma)$ is larger than the estimated values of r for the bulk of the sample period (for example, when $\gamma = .1$, $r_D(\gamma) = .349$).

IV. Conclusion and Further Thoughts

It seems intuitively plausible that the extent to which people will differ in their forecasts of inflation will increase with overall macro-economic uncertainty about the rate of inflation. This notion has been formalized here by presenting a theory of differential expectations within a rational expectations framework. The theory suggests that a positive relationship exists between the divergence of views about the future rate of inflation and the variance of aggregate demand shocks. Moreover, the particular model presented here suggests that if most of the changes in the variance of the rate of inflation over time are caused by changes in the underlying variance of aggregate demand, there will also be a positive association between the variability of the rate of inflation and the spread of expected inflation over different markets.

Those implications are tested empirically using directly measured price expectations from the Livingston and the *SRC* surveys. The empirical results support the implications of the model. That is, periods with large variances in the rate of inflation and large variances in the rate of change of nominal income also tend to be periods with large variances of inflationary expectations across survey respondents.

The results of this paper have a number of interesting implications which are worth brief mention. For example, Dennis Logue and Thomas Willett find the level and standard deviation of inflation to be positively correlated over time and across countries. This suggests in conjunction with the results of this paper that the level of inflation and the spread of expectations across individuals are positively related.

Along these lines, the survey data (see Wachtel) indicate that the level and variances of expected inflation are positively correlated. Greater variability of inflationary expectations also leads to greater incidence of unanticipated inflation and consequent unanticipated redistributions in nominally denominated asset markets. This may well be an important cost of high inflation.

Our results, therefore, give direct support to the view recently advanced by Gardner Ackley (p. 149), that the more variable a country's inflation, the less likely it is to have been anticipated. They also yield a justification for the use of the variance of inflationary expectations from survey data as a proxy for the variance of the inflationary process.[25]

Similarly, a positive association between the level of interest rates and the variability of expectations is expected. Nominal interest rates are higher because of the increased variance in the rate of inflation which in turn is caused by a higher variance of changes in aggregate demand. But since the variance of the rate of inflation and the variance of expectations from survey data are positively related, the last variance is actually a proxy for the previous variance, which explains its positive effect on nominal interest rates. Recognition of the nonuniformity of expectations in this regard suggests the need to reformulate Fisher's theory of interest for the case of heterogeneous expectations.[26]

Parks has shown that unanticipated inflation has a positive impact on the variance of relative price change assuming that expectations are uniform. The framework presented here may be used to extend his result partially to a world of nonuniform expectations. Such an attempt is made in our unpublished paper.

Appendix A: Exact Statement of Necessary Conditions for and Proof that $\bar{w} = \Sigma_\nu u(\nu)w(\nu)$ in (12) Converges in Probability to Zero

LEMMA: *Let $u(\nu) = f_\nu/N$ for all ν where*

$$N = \sum_{\nu=1}^{n} f_\nu > 0$$

and f_ν is nonnegative for all ν. Let $M = \max f_\nu$ be bounded from above. Let n be the total number of markets in the economy. Then

$$\lim_{n\to\infty} P_r[|\bar{w}| > 0] = 0$$

PROOF:

A direct consequence of Chebyshev's inequality is that a random variable converges in probability to its expected value if its variance tends to 0 as $n \to \infty$. It is therefore enough to show that the expected value and the variance of $\bar{w}$ both tend to zero as $n \to \infty$. But

$$Ew = \sum_{\nu=1}^{n} u(\nu)Ew(\nu) = 0$$

for any n. The variance of $\bar{w}$ is given by:

$$V(\bar{w}) = \sigma_w^2 \frac{\sum_{\nu=1}^{n} f_\nu^2}{(\sum_{\nu=1}^{n} f_\nu)^2} = \sigma_w^2 \frac{\sum_{\nu=1}^{n} f_\nu^2}{(nf_n^*)^2}$$

where $f_n^* = \dfrac{\sum_{\nu=1}^{n} f_\nu}{n}$

But since M is an upper bound on f_ν,

$$\text{(A1)} \quad V(\bar{w}) = \sigma_w^2 \frac{\sum_{\nu=1}^{n} f_\nu^2}{(nf_n^*)^2} < \sigma_w^2 \frac{M^2 n}{(nf_n^*)^2} = \sigma_w^2 \frac{M^2}{n(f_n^*)^2}$$

Since f_n^* is a simple average of f_ν, it is also bounded from above. Hence, as $n \to \infty$, the term on the extreme right-hand side of (A1) goes to zero, implying that $V(\bar{w}) \to 0$ as well.

Note that for the particular case in which all goods get the same weight, $u(\nu) = 1/n$ for all ν and the Lemma applies trivially since f_ν is obviously bounded from above. This is the case implicitly considered by Lucas.

Appendix B: Proof that the Rate of Change of Nominal Income Converges in Probability to Δx when the Number of Markets Becomes Large

Equation (4) implies that the total nominal income in period t is given by

$$\text{(A2)} \qquad NI_t = X_t \sum_{\nu=1}^{n} W(\nu)$$

where X_t and $W(\nu)$ are the antilogs of x_t and

[25]Such a proxy is used by Maurice Levi and John Makin in studies of the effects of inflation uncertainty on the Phillips curve.

[26]For such an attempt within the framework of a linear model, see Cukierman (1978).

$w(v)$, respectively. The average value of nominal income per market is

$$\text{(A3)} \qquad NI_t^* = \frac{X_t}{n}\sum_{v=1}^{n} W(v)$$

But $E[NI_t^* | X_t] = X_t EW$ and $Var\,[NI_t^* | X_t] = X_t^2\,(Var\,W/n)$. For $n \to \infty$ the variance of average nominal income per market tends to 0. It follows that NI_t^* conditioned on X_t converges in probability to $X_t EW$ whose rate of change is equal to $\partial \log X/\partial t = \Delta x$. The proof is completed by noting that the rates of change of aggregate and average nominal income are equal.

Appendix C: Proof of Proposition 2

Differentiating (23) partially with respect to σ_x^2 yields

$$\frac{\partial V(\pi^*)}{\partial \sigma_x^2} = 2\gamma^2 \frac{\theta^3(1-\theta)}{(1+\theta)^3}(\gamma\theta^2 + 2\theta - 1)$$

which will be positive if and only if the polynomial $p(\theta) = \gamma\theta^2 + 2\theta - 1 > 0$. The polynomial $p(\theta)$ has two roots, one positive (θ_1) and one negative, which is irrelevant, since θ is nonnegative. For nonnegative θ's, $p(\theta)$ increases in θ. Therefore, $p(\theta) > 0$ for any

$$\theta > \theta_1 = \frac{-1 + (1+\gamma)^{1/2}}{\gamma}$$

By rearranging this inequality and using the definition of θ, (15), and (17) it can be seen that this is equivalent to the condition $r < (1+\gamma)^{1/2}$ stated in the proposition.

Appendix D: Proof of Proposition 3

a) Follows from Proposition 2 and by differentiating $V(\pi)$ with respect to σ_x^2 which yields

$$\frac{\partial V(\pi)}{\partial \sigma_x^2} = \frac{2}{(1+\theta\gamma)^3}[1 + \theta\gamma + 2\theta\gamma(1-\theta)] > 0$$

b) Differentiating (23) and (25) with respect to σ_w^2 yields

$$\frac{\partial V(\pi)}{\partial \sigma_w^2} = -\frac{4(1-\theta)^2}{(1+\theta\gamma)^3} < 0$$

and

$$\frac{\partial V(\pi^*)}{\partial \sigma_w^2} = \gamma^2 \frac{(\theta(1-\theta))^2}{(1+\gamma\theta)^3}[3 + (\gamma - 4)\theta - 2\gamma\theta^2]$$

The last partial derivative will also be negative if and only if $p(\theta) = 2\gamma^2\theta^2 + (4-\gamma)\theta - 3 > 0$. This polynomial in θ has two roots, one of which is negative (and therefore irrelevant) and one which is positive and equal to:

$$\theta_1 = \frac{-(4-\gamma) + (\gamma^2 + 16(1+\gamma))^{1/2}}{4\gamma}$$

Furthermore, $\partial p(\theta)/\partial\theta = 4 + \gamma(4\theta - 1)$ which is always positive when $\gamma \le 4$. For $\gamma > 4$, *it is positive provided*

$$\text{(A4)} \qquad r < \frac{3\gamma + 4}{\gamma - 4}$$

Therefore, provided (A4) is fulfilled, $\partial V(\pi^*)/\partial\sigma_w^2$ will be negative when $\theta > \theta_1$. Rearranging, this last inequality becomes equivalent to the condition $r < r_D(\gamma)$. Since $r_D(\gamma) < (3\gamma + 4)/(\gamma - 4)$, (A4) is always fulfilled when $r < r_D(\gamma)$. This establishes sufficiency. When (A4) is violated, θ has to be smaller than θ_1 to assure that $\partial V(\pi^*)/\partial\sigma_w^2 < 0$, which is equivalent to the condition $r > r_D$, which must be satisfied when (A4) is violated since $(3\gamma + 4)/(\gamma - 4) > r_D$. This establishes necessity.

REFERENCES

G. Ackley, "The Costs of Inflation," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 149–54.

R. J. Barro, "Rational Expectations and the Role of Monetary Policy," *J. Monet. Econ.*, Jan. 1976, *2*, 1–32.

H. D. Brunk, *An Introduction to Mathematical Statistics*, 2d ed., Waltham; Toronto; London 1965.

J. Carlson, "Are Price Expectations Normally Distributed?," *J. Amer. Statist. Assn.*, Dec. 1975, *70*, 749–54.

———, "A Study of Price Forecasts," *Annals Econ. Soc. Measure.*, Winter 1977, *6*, 27–56.

A. **Cukierman**, "A Test of Expectational Processes Using Information from the Capital Market—The Israeli Case," *Int. Econ. Rev.*, Oct. 1977, *18*, 737–53.

———, "The Relationship Between Relative Prices and the General Price Level: A Suggested Interpretation," *Amer. Econ. Rev.*, June 1979, *69*, 444–47.

———, "Heterogeneous Inflationary Expectations, Fisher's Theory of Interest, and the Allocative Efficiency of the Bond Market," *Econ. Letters*, No. 2, 1978, *1*, 151–56.

——— **and P. Wachtel**, "Relative Price Variability and Non-Uniform Inflationary Expectations," unpublished paper, Aug. 1978.

F. T. **Juster and R. Comment**, "A Note on the Measurement of Price Expectations," unpublished paper, Inst. Soc. Res., Univ. Michigan 1978.

B. **Klein**, "Our New Monetary Standard: The Measurement and Effects of Price Uncertainty, 1880–1973," *Econ. Inquiry*, Dec. 1975, *13*, 461–83.

M. **Levi and J. Makin**, "Friedman and the Phillips Curve Stage II: Further Evidence," unpublished paper, Univ. British Columbia and Univ. Washington, Aug. 1977.

Joseph A. Livingston, surveys published twice yearly, *Philadelphia Sunday Bulletin*, 1948–71; *Philadelphia Inquirer*, 1972 on.

D. E. **Logue and T. D. Willett**, "A Note on the Relation Between the Rate and Variability of Inflation," *Economica*, May 1976, *46*, 151–58.

R. E. **Lucas, Jr.**, "Some International Evidence on Output-Inflation Tradeoffs," *Amer. Econ. Rev.*, June 1973, *63*, 326–34.

R. W. **Parks**, "Inflation and Relative Price Variability," *J. Polit. Econ.*, Feb. 1978, *86*, 79–96.

T. **Sargent**, "Rational Expectations, the Real Rate of Interest, and the Natural Rate of Unemployment," *Brookings Papers*, Washington 1973, *2*, 429–72.

D. **Vining and T. Elwertowski**, "The Relationship Between Relative Prices and the General Price Level," *Amer. Econ. Rev.*, Sept. 1976, *66*, 699–708.

P. **Wachtel**, "Surveys Measures of Expected Inflation and Their Potential Usefulness," in Joel Popkin, ed., *Analysis of Inflation: 1965–74*, Nat. Bur. Econ. Res. *Stud. in Income and Wealth*, Vol. 42, New York 1977.

Survey Research Center, (*SRC*) University of Michigan, *Quarterly Consumer Survey*, various years.

On the Mark: A Theory of Floating Exchange Rates Based on Real Interest Differentials

By JEFFREY A. FRANKEL*

Much of the recent work on floating exchange rates goes under the name of the "monetary" or "asset" view; the exchange rate is viewed as moving to equilibrate the international demand for stocks of assets, rather than the international demand for flows of goods as under the more traditional view. But within the asset view there are two very different approaches. These approaches have conflicting implications in particular for the relationship between the exchange rate and the interest rate.

The first approach might be called the "Chicago" theory because it assumes that prices are perfectly flexible.[1] As a consequence of the flexible-price assumption, changes in the nominal interest rate reflect changes in the expected inflation rate. When the domestic interest rate rises relative to the foreign interest rate, it is because the domestic currency is expected to lose value through inflation and depreciation. Demand for the domestic currency falls relative to the foreign currency, which causes it to depreciate instantly. This is a rise in the exchange rate, defined as the price of foreign currency. Thus we get a positive relationship between the exchange rate and the nominal interest differential.

The second approach might be called the "Keynesian" theory because it assumes that prices are sticky, at least in the short run.[2] As a consequence of the sticky-price assumption, changes in the nominal interest rate reflect changes in the tightness of monetary policy. When the domestic interest rate rises relative to the foreign rate it is because there has been a contraction in the domestic money supply relative to domestic money demand without a matching fall in prices. The higher interest rate at home than abroad attracts a capital inflow, which causes the domestic currency to appreciate instantly. Thus we get a *negative* relationship between the exchange rate and the nominal interest differential.

The Chicago theory is a realistic description when variation in the inflation differential is large, as in the German hyperinflation of the 1920's to which Frenkel first applied it. The Keynesian theory is a realistic description when variation in the inflation differential is small, as in the Canadian float against the United States in the 1950's to which Mundell first applied it. The problem is to develop a model that is a realistic description when variation in the inflation differential is moderate, as it has been among the major industrialized countries in the 1970's.

This paper develops a model which is a version of the asset view of the exchange rate, in that it emphasizes the role of expectations and rapid adjustment in capital markets. The innovation is that it combines the Keynesian assumption of sticky prices with the Chicago assumption that there are secular rates of inflation. It then turns out that the exchange rate is *negatively* related to the nominal interest differential, but *positively* related to the expected long-run inflation differential. The exchange rate differs from, or "overshoots," its equilibrium value by an amount

*Assistant professor, University of California-Berkeley. An earlier version of this paper was presented at the December 1977 meetings of the Econometric Society in New York. I would like to thank Rudiger Dornbusch, Stanley Fischer, Jerry Hausman, Dale Henderson, Franco Modigliani, and George Borts for comments.

[1]See papers by Jacob Frenkel and by John Bilson.

[2]The most elegant asset-view statement of the Keynesian approach is by Rudiger Dornbusch (1976c), to which the present paper owes much. Roots lie in J. Marcus Fleming and Robert Mundell (1964, 1968). They argued that if capital were perfectly mobile, a nonzero interest differential would attract a potentially infinite capital inflow, with a large effect on the exchange rate. More recently, Victor Argy and Michael Porter, Jürg Niehans, Dornbusch (1976a,b,c), Michael Mussa (1976) and Pentti Kouri (1976a,b) have introduced the role of expectations into the Mundell-Fleming framework.

which is proportional to the real interest differential, that is, the nominal interest differential minus the expected inflation differential. If the nominal interest differential is high because money is tight, then the exchange rate lies below its equilibrium value. But if the nominal interest differential is high merely because of a high expected inflation differential, then the exchange rate is equal to its equilibrium value, which over time increases at the rate of the inflation differential.

The theory yields an equation of exchange rate determination in which the spot rate is expressed as a function of the relative money supply, relative income level, the nominal interest differential (with the sign hypothesized negative), and the expected long-run inflation differential (with the sign hypothesized positive). The hypothesis is readily tested, using the mark/dollar rate, against the two alternative hypotheses: the Chicago theory, which implies a positive coefficient on the nominal interest differential, and the Keynesian theory, which implies a zero coefficient on the expected long-run inflation differential.

I. The Real Interest Differential Theory of Exchange Rate Determination

The theory starts with two fundamental assumptions. The first, interest rate parity, is associated with efficient markets in which the bonds of different countries are perfect substitutes:

$$d = r - r^* \tag{1}$$

where r is defined as the *log* of one plus the domestic rate of interest (which is numerically very close to the actual rate of interest for normal values) and r^* is defined as the *log* of one plus the foreign rate of interest.[3] If d is considered to be the forward discount, defined as the *log* of the forward rate minus the *log* of the current spot rate, then (1) is a statement of covered (or closed) interest parity. Under perfect capital mobility, that is, in the absence of capital controls and transactions costs, covered interest parity must hold exactly, since its failure would imply unexploited opportunities for certain profits.[4] However, d will be defined as the expected rate of depreciation; then (1) represents the stronger condition of uncovered (or open) interest parity. Of course if there is no uncertainty, as in a perfect foresight economy, then the forward discount is equal to the expected rate of depreciation, and (1) follows directly. If there is uncertainty and market participants are risk averse, then the assumption that there is no risk premium, though not precluded, is a strong one.[5]

The second fundamental assumption is that the expected rate of depreciation is a function of the gap between the current spot rate and an equilibrium rate, and of the expected long-run inflation differential between the domestic and foreign countries:

$$d = -\theta(e - \bar{e}) + \pi - \pi^* \tag{2}$$

where e is the *log* of the spot rate; π and π^* are the current rates of expected long-run inflation at home and abroad, respectively. (We can think of them as long-run rates of monetary growth that are known to the public.)[6] The *log* of the equilibrium exchange rate $\bar{e}$ is defined to increase at the rate $\pi - \pi^*$ in the absence of new disturbances; a more precise explanation will be given below. Equation (2) says that in the short run the exchange rate is expected to return to its

[3]The rates of interest referred to are instantaneous rates per unit of time, i.e., the forces of interest. In the absence of uncertainty (or, in the stochastic case, if the term structures of the interest differential and the forward discount contain no risk premium), equation (1) should also hold when the rates are defined over any finite interval τ, since the left-hand and right-hand sides would be equal to the expected future values of the left-hand side and right-hand side, respectively, of (1) integrated from 0 to τ.

[4]For evidence that the deviations from covered interest parity are smaller than transactions costs, see Frenkel and Richard Levich.

[5]My paper shows the conditions under which, despite risk aversion, there is a zero risk premium in the forward exchange rate. Briefly, the conditions are no outside nominal assets and no correlation between the values of currencies and the values of real assets.

[6]If there is no long-run growth of real income or technical change in the demand for money, then the rates of monetary growth are π and π^*. If there is long-run real growth or technical change, then we would adjust the monetary growth rates before arriving at the expected long-run inflation rates π and π^*.

equilibrium value at a rate which is proportional to the current gap, and that in the long run, when $e = \bar{e}$, it is expected to change at the long-run rate $\pi - \pi^*$. For the present, the justification for equation (2) will be simply that it is a reasonable form for expectations to take in an inflationary world. This claim will be substantiated in Appendix A, after a price-adjustment equation has been specified by a demonstration that (2), with a specific value implied for θ, follows from the assumptions of perfect foresight (or rational expectations in the stochastic case) and stability.[7] The rational value of θ will be seen to be closely related to the speed of adjustment in the goods market.

Combining equations (1) and (2) gives

$$(3) \qquad e - \bar{e} = -\frac{1}{\theta}[(r - \pi) - (r^* - \pi^*)]$$

We might describe the expression in brackets as the real interest differential.[8] Alternatively, note that in the long run when $e = \bar{e}$, we must have $\bar{r} - \bar{r}^* = \pi - \pi^*$, where $\bar{r}$ and $\bar{r}^*$ denote the long-run, short-term interest rates.[9] Thus the expression in brackets is equal to $[(r - r^*) - (\bar{r} - \bar{r}^*)]$, and the equation can be described intuitively as follows. When a tight domestic monetary policy causes the nominal interest differential to rise above its long-run level, an incipient capital inflow causes the value of the currency to rise proportionately above *its* equilibrium level.

For a complete equation of exchange rate determination, it remains only to explain $\bar{e}$. Assume that in the long run, purchasing power parity holds:

$$(4) \qquad \bar{e} = \bar{p} - \bar{p}^*$$

where $\bar{p}$ and $\bar{p}^*$ are defined as the *logs* of the equilibrium price levels at home and abroad, respectively.[10]

Assume also a conventional money demand equation:

$$(5) \qquad m = p + \phi y - \lambda r$$

where m, p, and y are defined as the *logs* of the domestic money supply, price level, and output. A similar equation holds abroad. Let us take the difference between the two equations:

$$(6) \qquad m - m^* = p - p^* + \phi(y - y^*) - \lambda(r - r^*)$$

Using bars to denote equilibrium values and remembering that in the long run, when $e = \bar{e}$, $\bar{r} - \bar{r}^* = \pi - \pi^*$, we obtain

$$(7) \qquad \bar{e} = \bar{p} - \bar{p}^* = \bar{m} - \bar{m}^* - \phi(\bar{y} - \bar{y}^*) + \lambda(\pi - \pi^*)$$

This equation illustrates the monetary theory of the exchange rate, according to which the exchange rate is determined by the relative supply of and demand for the two currencies. It says that in full equilibrium a given increase in the money supply inflates prices and thus raises the exchange rate proportionately, and that an increase in income or a fall in the expected rate of inflation raises the demand for money and thus lowers the exchange rate.

[7]Note, however, that the assumption of rational expectations or perfect foresight is not required for equation (2) and thus is not required for the model.

[8]This would not be quite right because the nominal interest rates are short term while the expected inflation rates are long term. However, it does turn out, with a price-adjustment equation such as is adopted in Appendix A, that $e - \bar{e}$ is proportional to the short-term real interest differential $[(r - Dp) - (r^* - Dp^*)]$.

[9]In words, long-run equality between the nominal interest differential and the expected inflation differential follows from interest rate parity (equality between the interest differential and expected depreciation) and long-run relative purchasing power parity (equality between depreciation and the inflation differential). An alternative argument is that in the long run international investment flows insure that real interest rates are equal across countries: $\bar{r} - \pi = \bar{r}^* - \pi^*$. The investment argument is not necessary, however, nor, if used, does it preclude the possibility of different real rates of interest in the short run, since even the most perfectly classical of economies have fixed capital stocks that earn nonzero profits in the short run.

[10]This assumption rules out the possibility of permanent shifts in the terms of trade and so gives essentially a one-good model. It does allow the possibility of temporary shifts; the existence of large departures from the Law of One Price for international trade even in closely matched categories of goods has been shown in studies by Peter Isard and by Irving Kravis and Robert Lipsey.

Substituting (7) into (3), and assuming that the current equilibrium money supplies and income levels are given by their current actual levels,[11] we obtain a complete equation of spot rate determination:

$$(8)\quad e = m - m^* - \phi(y - y^*) - \frac{1}{\theta}(r - r^*) + \left(\frac{1}{\theta} + \lambda\right)(\pi - \pi^*)$$

This is the equation which is tested empirically for the deutsche mark in Section III.

II. Testable Alternative Hypotheses

Equation (8) is reproduced here with an error term:[12]

$$(9)\quad e = m - m^* - \phi(y - y^*) + \alpha(r - r^*) + \beta(\pi - \pi^*) + u$$

where α $(= -1/\theta)$ is hypothesized negative and β $(= 1/\theta + \lambda)$ is hypothesized positive and greater than α in absolute value. Tests of a hypothesis are always more interesting if a plausible alternative hypothesis is specified. One obvious alternative hypothesis is Dornbusch's incarnation of the Keynesian approach, in which secular inflation is not a factor. In fact the model developed in this paper is the same as the Dornbusch model in the special case where $\pi - \pi^*$ always equals zero.[13] The testable hypothesis is $\beta = 0$.

Another—more conflicting—alternative hypothesis comes from the Chicago theory of the exchange rate attributable to Frenkel and Bilson. The variant presented by Bilson begins with a money demand equation like (5): $m - p = \phi y - \lambda r$. Subtracting the foreign version yields a relative money demand equation like (6): $(m - m^*) - (p - p^*) = \phi(y - y^*) - \lambda(r - r^*)$. Bilson then assumes that purchasing power parity always holds:

$$(10)\quad e = p - p^* = (m - m^*) - \phi(y - y^*) + \lambda(r - r^*)$$

An increase in the domestic interest rate lowers the demand for domestic currency and causes a depreciation. In terms of equation (9), α, the coefficient of the nominal interest differential, is hypothesized to be *positive* rather than negative.

The interest differential $(r - r^*)$ is viewed as representing the relative expected inflation rate $(\pi - \pi^*)$, either because international investment flows equate real rates of interest, or because interest rate parity insures that the interest differential equals expected depreciation, and purchasing power parity insures that depreciation equals relative inflation. Thus the expected inflation differential (were it directly observable) could be put into (10) instead of the nominal interest differential:

$$(11)\quad e = (m - m^*) - \phi(y - y^*) + \lambda(\pi - \pi^*)$$

In terms of equation (9), α is hypothesized to be zero and β to be positive, if we use a good proxy for $(\pi - \pi^*)$. Or, more generally, the hypothesis can be represented $\alpha + \beta = \lambda > 0$, $\alpha \geq 0$, $\beta \geq 0$. The relative size of α and β would depend on how good a proxy we have for the expected inflation differential.

Indeed Frenkel begins his analysis with the assumption of a Cagan-type money demand function, which uses the expected inflation rate rather than the interest rate:

$$(12)\quad m - p = \phi y - \lambda\pi$$

The assumption of purchasing power parity then gives equation (11) directly. Frenkel uses the expected rate of depreciation as reflected in the forward discount in place of the unobservable expected inflation differential which, in well-functioning bond markets,

[11] Actual money supplies and actual income levels can both easily be allowed to differ from their equilibrium levels, but these extensions of the analysis are omitted here in the interest of brevity.

[12] It is not stated here where the errors come from. Probably the most likely source of errors is the money demand equation (5), which would bias the coefficient of $(m - m^*)$ downward if it is not constrained to be one. This possibility is discussed below.

[13] See Dornbusch (1976c). The only other differences between the present model and the Dornbusch model are that the former is a two-country model, while the latter is a small-country model, and the latter uses a slightly different price-adjustment equation.

would be the same as using the nominal interest differential.

The argument that the nominal interest differential is equal to the expected inflation differential is the same as that given in the derivation of equation (7), and indeed equation (11) is identical to equation (7), except that (7) is hypothesized to hold only in long-run equilibrium while (10) is hypothesized to hold always.[14] The Frenkel-Bilson theory could be viewed as a special case of the real interest differential theory where the adjustment to equilibrium is assumed instantaneous; that is, θ is infinite, which of course is the same as α being zero.

The theory was originally tested by Frenkel on the German 1920–23 hyperinflation during which, it is argued, inflationary factors swamp everything else. In particular, variation in the expected inflation rate dwarfs variation in the real interest rate in the effect on the demand for money and thus the exchange rate. The argument is convincing; it is quite likely that the hypothesis $\alpha < 0$ would be rejected (or the hypothesis $\alpha \geq 0$ could not be rejected) if (9) were estimated on hyperinflation data. This just says that the Frenkel theory is the relevant one in the polar case when the inflation differential is very high and variable, much as the Dornbusch theory is clearly the relevant one in the polar case when the inflation differential is very low and stable.

It is the claim of the real interest differential theory that it is a realistic description in an environment of moderate inflation differentials such as has existed in the six years since the beginning of generalized floating in 1973, and that the alternative hypotheses break down in such an environment. Bilson has suggested and tested his theory for this period, and has claimed that empirically it works better than any alternative theory proposed.

The various alternative hypotheses are summarized in terms of equation (9):

Keynesian Model, Dornbusch (1976c):	$\alpha < 0$	$\beta = 0$
Chicago Model, Bilson:	$\alpha > 0$	$\beta = 0$
Frenkel:	$\alpha = 0$	$\beta > 0$
Real Interest Differential Model:	$\alpha < 0$	$\beta > 0$

III. Econometric Findings

In this section the real interest differential theory is tested on the mark/dollar exchange rate.[15] There are several good reasons for concentrating on this rate. The variation in the German-American inflation differential has been significant, as opposed to, for example, that in the Canadian-American or German-Swiss differentials. The exchange and capital markets are free from extensive government intervention in Germany and the United States, as opposed to, for example, those in the United Kingdom or Japan. In addition, the size of the German and American economies and the fact that there have been unexpectedly large upswings and downswings in the mark/dollar rate make this exchange rate the most important one to explain.

The sample used consisted of monthly observations between July 1974 and February 1978. The results were not greatly affected by the choice of monetary aggregate; only those using M_1 are reported in Table 1. Industrial production indices were used in place of national output, since the latter is not available on a monthly basis. Three-month money market rates were used for the nominal interest differential, divided by four to convert from a "percent per annum" basis to a three-month basis. Two kinds of proxies for the expected inflation differential were tried, both expressed on a three-month basis: past inflation differentials (averaged over the preceding year) and long-term interest differentials (under the rationale that the long-term real interest rates are equal).[16] The advantage

[14]In view of the result, (A4) in the Appendix, that the purchasing power parity gap is proportional to the real interest differential, it is not surprising that a theory which assumes that the former is always zero should also assume that the latter is always zero.

[15]Germany is viewed as the domestic country in the econometric equations.

[16]The equality of current long-term real rates of interest follows from the result that in the long run the short-term real rates are equal, and the requirement of a

TABLE 1—TEST OF REAL INTEREST DIFFERENTIAL HYPOTHESIS
(Sample: July 1974–February 1978)

Technique	Constant	$m - m_1^*$	$y - y^*$	$r - r^*$	$\pi - \pi^*$	R^2	D.W.	$\hat{\rho}$	Number of Observations
OLS	1.33	.87	−.72	−1.55	28.65	.80	.76		44
	(.10)	(.17)	(.22)	(1.94)	(2.70)				
CORC	.80	.31	−.33	−.259	7.72	.91		.98	43
	(.19)	(.25)	(.20)	(1.96)	(4.47)				
INST	1.39	.96	−.54	−4.75	27.42		1.00		42
	(.08)	(.14)	(.18)	(1.69)	(2.26)				
FAIR	1.39	.97	−.52	−5.40	29.40			.46	41
	(.12)	(.21)	(.22)	(2.04)	(3.33)				

Note: Standard errors are shown in parentheses.
Definitions: Dependent Variable (*log* of) Mark/Dollar Rate.
CORC = Iterated Cochrane-Orcutt.
INST = Instrumental variables for expected inflation differential are Consumer Price Index (*CPI*) inflation differential (average for past year), industrial Wholesale Price Index (*WPI*) inflation differential (average for past year), and long-term commercial bond rate differential.
FAIR = Instrumental variables are industrial *WPI* inflation differential and lagged values of the following: exchange rate, relative industrial production, short-term interest differential, and expected inflation differential. The method of including among the instruments lagged values of all endogenous and included exogenous variables, in order to insure consistency while correcting for first-order serial correlation, is attributed to Ray Fair
$m - m^*$ = *log* of German M_1/*U.S.* M_1
$y - y^*$ = *log* of German production/*U.S.* production
$r - r^*$ = Short-term German-*U.S.* interest differential
$(r - r^*)_{-1}$ = Short-term German-*U.S.* interest differential lagged
$\pi - \pi^*$ = Expected German-*U.S.* inflation differential, proxied by long-term government bond differential.

of the long-term interest differential is that it is capable of reflecting instantly the impact of new information such as the announcement of monetary growth targets. The long-term government bond rate differential is the proxy used in the reported regressions, though other proxies are used as instrumental variables. Details on the data are given in Appendix B.

In each regression the signs of all coefficients are as hypothesized under the real interest differential model. When the single equation estimation techniques are used, the significance levels are weak, especially when iterated Cochrane-Orcutt is used to correct for high first-order autocorrelation.

But when instrumental variables are used to correct for the shortcomings of the expected inflation proxy, the results improve markedly. The coefficient on the nominal interest differential is significantly less than zero. This result is all the more striking when it is kept in mind that the null hypothesis of a zero or positive coefficient is a plausible and seriously maintained hypothesis; the Chicago (Frenkel-Bilson) hypothesis is rejected in this data sample. The coefficient on the expected long-run inflation differential is significantly greater than zero. Thus the unmodified Keynesian (Dornbusch) hypothesis is also rejected. Furthermore, as predicted by the real interest differential model the coefficient on the expected long-run inflation differential is significantly greater than the absolute value of the coefficient on the nominal interest differential.

Several other points are also notably supportive of the theory. (I concentrate on the last regression in Table 1.) The coefficient of the relative money supply is not only significantly positive, but is also insignificantly less than 1.0. The coefficient of relative production is significantly negative, and its point estimate of approximately −.5 suits well its interpretation as the elasticity of money demand with respect to income. The sum of

rational term structure that the long-term real interest differential be the average of the expected short-term real interest differentials.

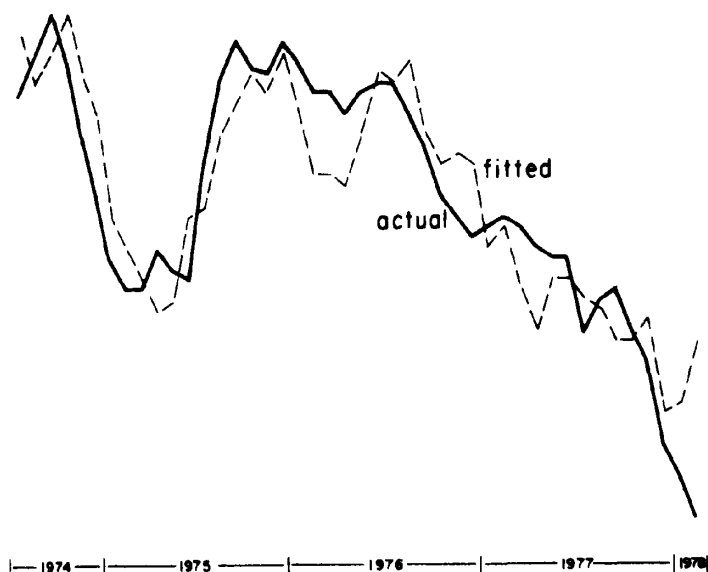



FIGURE 1. PLOT OF (*log* OF) MARK/DOLLAR RATE, *OLS* REGRESSION FROM TABLE 1

the (negative) coefficient on the nominal interest differential and the coefficient on the expected inflation differential is an estimate of the semielasticity of money demand with respect to the interest rate; when converted to a per annum basis, the estimate is 6.0, which provides another favorable cross-check.[17]

The point estimate of α is -5.4. This implies that when a disturbance creates a deviation from purchasing power parity, $(1 - 1/5.4 =)$ 81.5 percent of the deviation is expected to remain after one quarter, and $(.815^4 =)$ 44.1 percent is expected to remain after one year. The estimate of θ on a per annum basis is $(-log\ .441 =)$ .819. Previous work on the speed of adjustment to purchasing power parity is less definitive than estimates of money demand elasticities, but the present estimates of the expected speed of adjustment appear reasonable.[18]

As a final indication of the support Table 1 provides for the real interest differential hypothesis, the R^2s are high. Figure 1 shows a plot of the equation's predicted values and the actual exchange rate values. The equation tracks the mark's 1974 appreciation, 1975 depreciation, and 1976–77 appreciation.[19]

To apply the estimated equation, let us convert it to the form:

$$e = 1.39 + (m - m^*) - .52(y - y^*) - 1.35(r - r^*) + 7.35(\pi - \pi^*)$$

where α and β have been divided by four for use with per annum interest rates and the coefficient on the relative money supply has been set to 1.0. The expression can be decomposed into the equilibrium exchange rate

$$\bar{e} = 1.39 + (m - m^*) - .52(y - y^*) + 6.00(\pi - \pi^*)$$

[17]The semielasticity estimate and an average interest rate of around 6 percent imply an interest elasticity of around (6.0 x .06 =) .36, which is in the range of estimates of the long-run elasticity made by Stephen Goldfeld and others.

[18]Hans Genberg estimates for Germany that 37 percent of an initial divergence from purchasing power parity disappears after one year.

[19]The equation fails to track the continued sharp depreciation of the dollar in January and February of 1978. The regressions that were reported in earlier versions of this paper did not include this period, and consequently appeared more favorable to the real interest differential model.

and the size of the overshooting

$$e - \bar{e} = -1.35[(r - \pi) - (r^* - \pi^*)]$$

As an illustration, let us conduct the hypothetical experiment of an unexpected 1 percent expansion in the *U.S.* relative money supply. If the monetary expansion is considered a once-and-for-all change, then the equilibrium mark/dollar rate decreases by 1.0 percent. But in the short run the expansion also has liquidity effects; the interest semi-elasticity of 6.00 implies a fall in the nominal interest rate of (1 percent/6.00 =) 17 basis points.[20] This fall in the real interest differential induces an incipient capital outflow which in turn causes the currency to depreciate further, until it overshoots its new equilibrium by (1.35 x .17 percent =) .23 percent. The total initial depreciation is 1.23 percent.

This calculation assumes no change in the expected inflation rate. If the monetary expansion signals a new higher target for monetary growth, then the effect could be much greater.[21] Suppose the annualized 12 percent increase raises the expected inflation rate by, say, 1 percent per annum. Then there will be an additional depreciation of 6.00 percent on account of the lower demand for money in long-run equilibrium plus 1.35 percent more overshooting on account of the further reduced real interest differential. Thus the *total* initial depreciation would be 8.58 percent, of which 7.00 percent represents long-run equilibrium and 1.58 percent represents short-run overshooting.

After the initial effects, the system moves toward the new equilibrium as described in Appendix A, provided capital is perfectly mobile and future money supplies do not deviate from their expected values. American goods are cheaper than German goods; higher demand will gradually drive up American prices faster than the rate of monetary growth, which in turn will drive up *U.S.* nominal interest rates, reduce the overshooting, and cause the spot rate to rise back towards its new equilibrium. After a year, approximately 44 percent of the initial real interest differential and purchasing power parity deviation will have been closed. In the meantime, there should be an expansionary effect on demand for *U.S.* output; lower real *U.S.* prices will stimulate net exports and lower real *U.S.* interest rates will stimulate investment. However, any effects on output have not been modelled in this paper.[22]

IV. Econometric Extensions

It is possible that adjustment in capital markets to changes in the interest differential is not instantaneous, and that lagged interest differentials should be included in the regressions. Formally, we could argue that due to transactions costs, the forward discount adjusts fully to the interest differential with a one-month lag:

$$(13) \quad d = h(r - r^*) + (1 - h)(r - r^*)_{-1}$$

When (13) is used in place of (1), the spot rate equation (8) is replaced by

$$(14) \quad e = (m - m^*) - \phi(y - y^*) - (h/\theta)(r - r^*) - (1 - h)(1/\theta)(r - r^*)_{-1} + ((1/\theta) + \lambda)(\pi - \pi^*)$$

The results of regressions with a lagged interest differential are reported in Table 2. The coefficient on the lagged interest differential is insignificantly less than zero. This evidence supports the idea that capital is perfectly mobile.

There are several reasons why one might wish to constrain the coefficient on the relative money supply to be 1.0 in these regressions, in effect moving the relative money

[20] Here we are assuming (for the first time) that the estimated long-run interest semielasticity holds in the short run as well. If money demand is subject to lagged adjustment, then the short-run effect of a monetary contraction on the interest rate and hence on the exchange rate would be *greater* than that calculated here. However the theory and econometrics behind equation (8) would be completely unaffected.

[21] On the other hand, if some of the expansion is expected to be reversed in the following month, then the decrease in the equilibrium rate would be correspondingly *smaller*.

[22] The assumption of exogenous output can be relaxed by assuming that output is demand determined. Then a necessary condition for overshooting is that the elasticity of demand with respect to relative prices is less than 1.0. See the Appendix to Dornbusch (1976c).

TABLE 2—TEST WITH LAGGED INTEREST DIFFERENTIAL
(Sample: July 1974–February 1978)

Technique	Constant	$m - m^*$	$y - y^*$	$r - r^*$	$(r - r^*)_{-1}$	$\pi - \pi^*$	R^2	$D.W.$	$\hat{\rho}$	Number of Observations
OLS	1.34	.90	−.76	−3.34	2.27	29.33	.80	.82		44
	(.11)	(.18)	(.23)	(3.09)	(3.04)	(2.87)				
CORC	.56	.22	−.30	−3.00	−2.92	7.14	.92		.99	43
	(.29)	(.25)	(.20)	(1.90)	(1.73)	(4.32)				
INST	1.39	.96	−.52	−4 11	−.90	27 10		.97		42
	(.09)	(.15)	(.20)	(2.48)	(2.54)	(2.48)				
FAIR	1.04	.31	−.16	−6.83	−3.37	29.87			.81	43
	(.19)	(.33)	(.24)	(2.63)	(2 30)	(6 59)				

Note: Standard errors are shown in parentheses. Dependent Variable. (*log* of) Mark/Dollar Rate. See Table 1 for definitions.

supply variable to the left-hand side of the equation. First, our a priori faith in a unit coefficient is very high. It is hard to believe that the system could fail in the long run to be homogeneous of degree zero in the exchange rate and relative money supply. Second, errors in the money demand equation are known to have been large over the last few years. Such errors, since they are correlated with the money stock variable, would bias the coefficient downward; indeed, the coefficient estimate in one regression in Table 1 appears significantly less than 1.0. By constraining the money supply coefficient to be 1.0, we make sure that any possible errors in the money demand equations will go into the dependent variable, that is, will be uncorrelated with any of the independent variables; thus they cannot bias the coefficients on the interest and expected inflation differentials, which are our primary objects of concern. A third reason for constraining the coefficient is to remove the simultaneity problem which otherwise occurs if central banks vary their money supplies in response to the exchange rate. The argument even extends to direct exchange market intervention, which has been prevalent under managed floating. The right-hand side variables determine relative money demand; changes in money demand can be reflected in either money supplies (the monetary approach to the balance of payments) or the exchange rate (the monetary approach to the exchange rate), depending on government intervention policy.

Table 3 reports the constrained regressions. The results are very similar to those in Table 1. The R^2s indicate that over 90 percent of the variation in the dependent variable is explained; the remaining 10 percent could be attributed to errors in the two countries' money demand equations.

TABLE 3—CONSTRAINED COEFFICIENT ON RELATIVE MONEY SUPPLIES
(Sample: July 1974–February 1978)

Technique	Constant	$y - y^*$	$r - r^*$	$\pi - \pi^*$	R^2	$D W$	$\hat{\rho}$
OLS	1.40	−.69	−1.77	30 17	.92	.79	
	(.02)	(.21)	(1.91)	(1 68)			
CORC	1.16	−.41	−1 55	10.13	.96		.98
	(.13)	(.22)	(2.11)	(4.82)			
INST	1.41	−.52	−4.84	27.73		1.01	
	(.01)	(18)	(1.64)	(1.47)			
FAIR	1.43	−.31	−5.61	34.01			69
	(.03)	(.26)	(2.70)	(4.24)			

Note: Standard errors are shown in parentheses. Dependent Variable: (*log* of) Mark/Dollar Rate ÷ German M_1/*U.S.* M_1. See Table 1 for definitions.

V. Summary

The model developed in this paper is a version of the asset view of the exchange rate, in that it emphasizes the role of expectations and rapid adjustment in capital markets. It shares with the Frenkel-Bilson (Chicago) model an attention to long-run monetary equilibrium. A monetary expansion causes a long-run depreciation because it is an increase in the supply of the currency, and an increase in expected inflation causes a long-run depreciation because it decreases the demand for the currency.

On the other hand, the model shares with the Dornbusch (Keynesian) model the assumption that sticky prices in goods markets create a difference between the short run and the long run. When the nominal interest rate is low relative to the expected inflation rate, the domestic economy is highly liquid. An incipient capital outflow will cause the currency to depreciate, until there is sufficient expectation of future appreciation to offset the low interest rate. The exchange rate overshoots its equilibrium value by an amount proportional to the real interest differential.

The real interest differential model includes both the Frenkel-Bilson and Dornbusch models as polar special cases. When the spot rate equation (8) is econometrically estimated for the mark/dollar rate from July 1974 to February 1978, the evidence clearly supports the model against the two alternatives.

Appendix A: The Price Equation and the Path to Equilibrium

In this appendix we examine the consequences of an additional assumption, a price equation. Unless there is some stickiness in p, it cannot differ from $\bar{p}$ and thus the domestic real interest rate cannot differ from the foreign real interest rate, or the exchange rate from the relative price level. This stickiness can be embodied in the assumption that prices are fixed at a moment in time, but move gradually toward equilibrium. In an environment of secular monetary growth, it is necessary that when prices reach their equilibrium, they are increasing at the secular rate. The simplest possible price equation meeting these requirements is

$$\text{(A1)} \qquad Dp = \delta(e - p + p^*) + \pi$$

This equation can be rationalized by expressing the rate of change of prices as the sum of a mark-up term π, representing the pass-through of domestic cost inflation and an excess demand adjustment term, where excess demand is assumed a function of the purchasing power parity gap $(e - p + p^*)$.[23] Assuming that the analogous equation holds abroad, the relative price level changes according to

$$\text{(A2)} \quad D(p - p^*) = \delta(e - p + p^*) + \pi - \pi^*$$

where δ has been redefined to be the sum of the domestic and foreign adjustment parameters.

The purchasing power parity gap (also called the real exchange rate) can be shown to be proportional to the real interest differential. Substituting (6) into (7) implies

$$\bar{e} = p - p^* - \lambda[(r - \pi) - (r^* - \pi^*)]$$

which with (3) implies

$$\text{(A3)} \quad e - p + p^* = -\left(\frac{1}{\theta} + \lambda\right)[(r - \pi) - (r^* - \pi^*)]$$

Now we use (A2) to solve out $(\pi - \pi^*)$, and collect terms to arrive at the promised result:

$$\text{(A4)} \quad e - p + p^* = -\frac{1 + \lambda\theta}{\theta - (1 + \lambda\theta)\delta}[(r - Dp) - (r^* - Dp^*)]$$

Let us now proceed to derive the path from the initial point after a disturbance (short-run equilibrium) to long-run equilibrium. We already know from equations (3) and (A3) that the gap between e and its equilibrium and the purchasing power parity gap are each proportional to $[(r - \pi) - (r^* - \pi^*)]$, so they must be proportional to each other:

[23] The rationalization necessarily implies a disequilibrium formulation of the goods market. More sophisticated price equations give very similar results, but are not considered here.

(A5) $(e - p + p^*) = (1 + \lambda\theta)(e - \bar{e})$

Using $\bar{e} - \bar{p} + \bar{p}^* = 0$ and (A5),

(A6) $$e - p + p^* = (e - \bar{e}) - (p - \bar{p}) + (p^* - \bar{p}^*)$$
$$= -\frac{1 + \lambda\theta}{\lambda\theta}[(p - p^*) - (\bar{p} - \bar{p}^*)]$$

Substituting (A6) into (A2),

(A7) $$D(p - p^*) = -\delta(1 + \lambda\theta)/\lambda\theta\,[(p - p^*) - (\bar{p} - \bar{p}^*)] + \pi - \pi^*$$

This differential equation has the solution

(A8) $$(p - p^*)_t = (\bar{p} - \bar{p}^*)_t + \exp[-\delta(1 + \lambda\theta/\lambda\theta)t][(p - p^*)_0 - (\bar{p} - \bar{p}^*)_0]$$

The relative price level moves toward its equilibrium at a speed which is proportional to the gap. The equilibrium relative price level, it must be remembered, is itself increasing at the rate $\pi - \pi^*$.

An analogous equation holds for e. Equations (A5) and (A6) tell us

(A9) $$e - \bar{e} = -\frac{1}{\lambda\theta}[(p - p^*) - (\bar{p} - \bar{p}^*)]$$

Taking the time derivative,

(A10) $$De = -\frac{1}{\lambda\theta}D[(p - p^*) - (\bar{p} - \bar{p}^*)] + D\bar{e}$$
$$= -\frac{\delta(1 + \lambda\theta)}{\lambda\theta} \cdot (e - \bar{e}) + \pi - \pi^*$$

This differential equation has the solution

(A11) $$e_t = \bar{e}_t + \exp[-(\delta(1 + \lambda\theta)/\lambda\theta)t](e - \bar{e})_0$$

Comparing (A10), the expression for the rate of change of the spot rate if there are no further disturbances, with (2), the expression for the expected rate of change of the exchange rate, we see that the two are of the same form. Perfect foresight (or rational expectations in the stochastic case) holds if $\theta = \delta(1 + \lambda\theta)/\lambda\theta$, which has the solution

(A12) $$\tilde{\theta}_1 = \frac{\delta\lambda + ((\delta\lambda)^2 + 4\delta\lambda)^{1/2}}{2\lambda}$$
$$\tilde{\theta}_2 = \frac{\delta\lambda - ((\delta\lambda)^2 + 4\delta\lambda)^{1/2}}{2\lambda}$$

Here we throw out the negative root because θ was assumed positive when (2) was specified. We can see that $\tilde{\theta}_1$ increases with δ, the speed of adjustment in goods markets. In turn, we know from equation (3) that the sensitivity of the exchange rate to monetary changes decreases with θ. The implication is that the slower is adjustment in the goods market, the more volatile must the exchange rate be in order to compensate.

It is easy to show that we could have derived (2) from the rest of the model and the assumptions of perfect foresight and stability, instead of assuming the form of expectations directly. Substituting the relative money demand equation (6) into the interest parity condition (1),

(A13) $$d = (1/\lambda)[(p - p^*) - (m - m^*) + \phi(y - y^*)]$$
$$= (1/\lambda)[(p - p^*) - (\bar{p} - \bar{p}^*)] + \pi - \pi^*$$

The perfect foresight assumption is $d = De$. Equation (A13) and the price equation (A2) can be represented in matrix form:

$$\begin{bmatrix} De \\ D(p - p^*) \end{bmatrix} = \begin{bmatrix} 0 & 1/\lambda \\ \delta & -\delta \end{bmatrix} \begin{bmatrix} e \\ (p - p^*) \end{bmatrix} + \begin{bmatrix} -(1/\lambda)(\bar{p} - \bar{p}^*) + \pi - \pi^* \\ \pi - \pi^* \end{bmatrix}$$

Let $-\theta_1$ and $-\theta_2$ be the characteristic roots:

$$\begin{vmatrix} \theta & 1/\lambda \\ \delta & -\delta + \theta \end{vmatrix} = -\delta\theta + \theta^2 - \delta/\lambda = 0$$

The solution is given by (A12). The path of e is given by

$$(e - \bar{e})_t = a_1 \exp(-\theta_1 t) + a_2 \exp(-\theta_2 t)$$

The system is stable if and only if $a_2 = 0$, which, with the initial condition $a_1 = (e - \bar{e})_0$, implies equation (2), and the positive root from (A12).

APPENDIX B

The data are described as follows:

Spot rate: Monthly averages of Dollars per Mark, *Federal Reserve Bulletin* (*FRB*).

Money supply: Germany: Position at end of month, seasonally adjusted, in billions of marks, Deutsche Bundesbank (*DB*). United States: Averages of daily figures, seasonally adjusted, in billions of dollars, from the *Economic Report of the President* (*ERP*), *FRB*, and *Economic Indicators* (*EC*).

Industrial production: Germany: Seasonally adjusted, *DB*. United States: Seasonally adjusted, *ERP* and *Statistical Releases*.

Short-run interest rates: Representative bond equivalent yields on major 3–4-month money market instruments, excluding Treasury Bills, *World Financial Markets* (*WFM*).

Expected inflation rates: Long-term government bond yields: at or near end of month, *WFM*. Long-term commercial bond yields: at or near end of month, *WFM*.

Wholesale price index (average logarithmic rate of change over preceding year): Germany: Industrial products, seasonally adjusted, *DB*. United States: Industrial, seasonally adjusted, *ERP*, *FRB*, *International Financial Statistics* (*IFS*), and *Business Week*.

Consumer price index (average logarithmic rate of change over preceding year): Germany: Cost of living, seasonally adjusted, *DB*. United States: Urban dwellers and clerical workers, *ERP*, *IFS*, and *EC*.

The three-month interest rates were converted from percent per annum terms to the absolute three-month rate of return (the theoretically correct variable) by dividing by 400. The expected inflation rates were also converted to a per quarter basis for consistency.

REFERENCES

V. Argy and M. Porter, "The Forward Exchange Market and the Effects of Domestic and External Disturbances Under Alternative Exchange Rate Systems," *Int. Monet. Fund Staff Papers*, Nov. 1972, *19*, 503–32.

J. Bilson, "The Monetary Approach to the Exchange Rate—Some Empirical Evidence," *Int. Monet. Fund Staff Papers*, Mar. 1978, *25*, 48–75.

R. Dornbusch, (1976a) "The Theory of Flexible Exchange Rates and Macroeconomic Policy," *Scand. J. Econ.*, June 1976, *78*, 255–75.

———, (1976b) "Exchange Rate Expectations and Monetary Policy," *J. Int. Econ.*, Aug. 1976, *6*, 231–44.

———, (1976c) "Expectations and Exchange Rate Dynamics," *J. Polit. Econ.*, Dec. 1976, *84*, 1161–76.

——— **and P. Krugman**, "Flexible Exchange Rates in the Short Run," *Brookings Papers*, Washington 1976, *3*, 537–84.

J. M. Fleming, "Domestic Financial Policies Under Fixed and Floating Exchange Rates," *Int. Monet. Fund Staff Papers*, Nov. 1962, *9*, 369–79.

J. A. Frankel, "The Diversifiability of Exchange Risk," *J. Int. Econ.*, Aug 1979, *9*.

J. Frenkel, "A Monetary Approach to the Exchange Rate: Doctrinal Aspects and Empirical Evidence," *Scand. J. Econ.*, June 1976, *78*, 255–76.

——— **and R. Levich**, "Covered Interest Arbitrage: Unexploited Profits?," *J. Polit. Econ.*, Apr. 1975, *83*, 325–29.

H. Genberg, "Purchasing Power Parity Under Fixed and Flexible Exchange Rates," *J. Int. Econ.*, May 1978, *8*, 247–76.

S. Goldfeld, "The Case of the Missing Money," *Brookings Papers*, Washington 1976, *3*, 683–730.

D. Henderson, "Monetary, Fiscal and Exchange Rate Policy in a Two-Country Short-Run Macroeconomic Model," mimeo., Federal Reserve Board, Washington 1975.

P. Isard, "How Far Can We Push the 'Law of One Price'?," *Amer. Econ. Rev.*, Dec. 1977, *67*, 942–48.

P. Kouri, (1976a) "The Exchange Rate and the Balance of Payments in the Short Run and in the Long Run," *Scand. J. Econ.*, June 1976, *78*, 280–304.

———, (1976b) "Foreign Exchange Market Speculation and Stabilization Policy Under Flexible Exchange Rates," paper presented at Conference on the Political Economy of Inflation and Unemployment in Open Economies, Athens, Oct. 1976.

I. Kravis and R. Lipsey, "Export Prices and the Transmission of Inflation," *Amer. Econ. Rev. Proc.*, Feb. 1977, *67*, 155–63.

Robert Mundell, "Exchange Rate Margins and Economic Policy," in J. Carter Murphy, ed., *Money in the International Order*, Dallas 1964.

———, *International Economics*, New York 1968.

M. Mussa, "The Exchange Rate, the Balance of Payments and Monetary and Fiscal Policy Under a Regime of Controlled Floating," *Scand. J. Econ.*, June 1976, *78*, 229–48.

———, "Real and Monetary Factors in a Dynamic Theory of Foreign Exchange," mimeo., Univ. Chicago, Apr. 1977.

J. Niehans, "Some Doubts About the Efficacy of Monetary Policy Under Flexible Exchange Rates," *J. Int. Econ.*, Aug. 1975, *5*, 275–81.

M. v. N. Whitman, "Global Monetarism and the Monetary Approach to the Balance of Payments," *Brookings Papers*, Washington 1975, *3*, 491–556.

Board of Governors of the Federal Reserve System, *Statistical Releases*, Washington, various issues.

———, *Fed. Res. Bull.*, Washington, various issues.

Business Week, various issues.

Deutsche Bundesbank, Statistical Supplements to Monthly Reports, Bonn, various issues.

International Monetary Fund, *International Financial Statistics*, Washington, various issues.

Morgan Guaranty and Trust Co., *World Financial Markets*, New York, various issues.

U.S. Congress, Joint Economy Committee, *Economic Indicators*, Washington, various issues.

U.S. Council of Economic Advisors, *Economic Report of the President*, Washington, various issues.

Economic Theory of Choice and the Preference Reversal Phenomenon

By DAVID M. GRETHER AND CHARLES R. PLOTT*

A body of data and theory has been developing within psychology which should be of interest to economists. Taken at face value the data are simply inconsistent with preference theory and have broad implications about research priorities within economics. The inconsistency is deeper than the mere lack of transitivity or even stochastic transitivity. It suggests that no optimization principles of any sort lie behind even the simplest of human choices and that the uniformities in human choice behavior which lie behind market behavior may result from principles which are of a completely different sort from those generally accepted. This paper reports the results of a series of experiments designed to discredit the psychologists' works as applied to economics.

The phenomenon is characterized by the following stylized example. Individuals under suitable laboratory conditions are asked if they prefer lottery A to lottery B as shown in Figure 1. In lottery A a random dart is thrown to the interior of the circle. If it hits the line, the subject is paid $0 and if it hits anywhere else, the subject is paid $4. Notice that there is a very high probability of winning so this lottery is called the P bet, standing for probability bet. If lottery B is chosen, a random dart is thrown to the interior of the circle and the subject receives either $16 or $0 depending upon where the dart hits. Lottery B is called the $ bet since there is a very high maximum reward. After indicating a preference between the two lotteries, subjects are asked to place a monetary value on each of the lotteries.

*Professors of economics, California Institute of Technology. The financial support supplied by the National Science Foundation and the National Aeronautics and Space Administration is gratefully acknowledged. We wish to express our appreciation to Brian Binger, Elizabeth Hoffman, and Steven Matthews, who served as research assistants.

Psychologists have observed that a large proportion of people will indicate a preference for lottery A, the P bet, but place a higher value on the *other* lottery, the $ bet. The following argument will help us see one way in which this behavior violates preference theory. Let w = initial wealth; $(z,1,0)$ = the state in which A is held and the wealth level is z; $(z,0,1)$ = the state in which B is held and the wealth level is z; $(z,0,0)$ = the state in which neither A nor B are held and the wealth level is z; $\$(A)$ and $\$(B)$ are the respective selling limit prices; $\sim$ and $\succ$ indicate indifference and preference, respectively.

(1) $(w+\$(A),0,0) \sim (w,1,0)$ by definition of $\$(A)$

(2) $(w+\$(B),0,0) \sim (w,0,1)$ by definition of $\$(B)$

(3) $(w,1,0) \succ (w,0,1)$ by the statement of preference of A over B

(4) $(w+\$(A),0,0) \succ (w+\$(B),0,0)$ by transitivity

(5) $\$(A) > \$(B)$ by positive "utility value" of wealth

Though (5) follows from the theory, it is not observed.

Notice this behavior is not simply a violation of some type of expected utility hypothesis. The preference measured one way is the *reverse* of preference measured another and seemingly theoretically compatible way. If indeed preferences exist and if the principle of optimization is applicable, then an individual should place a higher reservation price on the object he prefers. The behavior as observed appears to be simply inconsistent with this basic theoretical proposition.

If the results are accepted uncritically and extended to economics, many questions are raised. If preference theory is subject to systematic exception in these simple cases, how many other cases exist? What type of theory of choice can serve as a basis for market theory and simultaneously account for

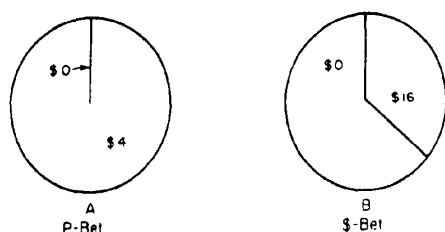



FIGURE 1

these data? Could such an alternative theory also serve as a basis for welfare economics? Should special extensions of the theory of market choice to other situations such as crime (Gary Becker, 1968), suicide (Daniel Hammermesh and Neal Soss), marriage (Becker, 1973, 1974), extramarital affairs (Ray Fair), politics, etc. be called into question? How are we to regard cost-benefit measures once we have accepted the fact that the sign of the benefit-minus-cost figure can be *reversed* by simply measuring preference in terms of "most preferred" options rather than in terms of a limit selling or purchase price?

There is little doubt that psychologists have uncovered a systematic and interesting aspect of human choice behavior. The key question is, of course, whether this behavior should be of interest to economists. Specifically it seems necessary to answer the following: 1) Does the phenomenon exist in situations where economic theory is generally applied? 2) Can the phenomenon be explained by applying standard economic theory or some immediate variant thereof?

This study was designed to answer these two questions. The experiments prior to those reported here were not designed with economics audiences in mind and thus do not answer either question though they are suggestive. In the first section we review the earlier experiments and their shortcomings from an economics point of view. Our experimental design is covered in the second section, and our results are reviewed in the third section. In the end we will conclude that the answer to the first question is "yes" and the answer to the second appears to be "no." As reflected in our concluding remarks, we remain as perplexed as the reader who has just been introduced to the problem.

I. Existing Experimental Work

Experimental results of direct relevance are reported in four papers (Sarah Lichtenstein and Paul Slovic, 1971, 1973; Harold Lindman; Slovic). The experiments are listed in Table 1 beside an array of theories, each of which would either render the experiment irrelevant from an economist's point of view or explain the results in terms of accepted theory. Along with the economic-theoretic explanations, we have listed some psychological-theoretic explanations and some theories which seek to explain the results as artifacts of experimental methods.

A. *Economic-Theoretic Hypotheses*

THEORY 1: *Misspecified Incentives*. Almost all economic theory is applied to situations where the agent choosing is seriously concerned or is at least choosing from among options that in some sense matter. No attempt is made to expand the theory to cover choices from options which yield consequences of no importance. Theories about decision-making costs do suggest that unmotivated choice behavior may be very different from highly motivated choice behavior, but the differences remain essentially unexplored. Thus the results of experiments where subjects may be bored, playing games, or otherwise not motivated, present no immediate challenges to theory.

On this basis several experiments can be disregarded as applying to economics even though they may be very enlightening for psychology. In Lichtenstein and Slovic (1971) the first two experiments were made from gambles which involved imaginary money and in the third, the gambles were for points the value of which were not revealed until after the choices were made. All experiments in Lindman involved gambles for "imaginary money." Three of the experiments of Slovic dealt with the choices among fictitious commodity bundles. The only experiments which cannot be criticized on this ground are

TABLE 1—COEXISTING EXPERIMENTAL RESULTS: RELEVANCE AND POSSIBLE EXPLANATIONS

	Theoretical Criticism and/or Explanation	Lichtenstein & Slovic (1971) Experiment			Lichtenstein & Slovic (1973)	Lindman (1971)	Slovic (1975) Experiment				This Study Experiment	
		1	2	3			1	2	3	4	1	2
	Economic Theory											
1.	Misspecified Incentives	I	I	I	N	I	I	I	N	I	N	N
2.	Income Effects	N	N	E	?	N	N	N	E	N	N	N
3.	Indifference	N	I	I	I	I	I	I	I	I	N	N
	Psychological											
4.	Strategic Responses	E	E	E	E	E	N	N	N	N	E	N
5.	Probabilities	I	I	N	?	N	N	N	N	N	N	N
6.	Elimination by Aspect	N	N	N	N							N
7.	Lexicographic Semiorder	N	N	N	N							N
8.	Information Processing: Decision Costs	E	E	E	?	E	E	E	E	E	N	N
9.	Information Processing: Response Mode, Easy Justification	E	E	E	E	E	E	E	E	E	E	E
	Experimental Methods											
10.	Confusion and Misunderstanding	N	N	N	N	N	N	N	N	N	N	N
11.	Frequency Low	N	N	N	N	N	N	N	N	N	N	N
12.	Unsophisticated Subjects	?	?	?	N	?	?	?	?	N	N	N
13.	Experimenters Were Psychologists	I	I	I	I	I	I	I	I	I	N	N

I = The experiment is irrelevant to economics because of the reason or theory.

N = The experimental results cannot be explained by this reason or theory.

E = The experimental results are consistent with the reason or theory.

? = Data insufficient.

Lichtenstein and Slovic (1973), which was conducted on a Las Vegas casino floor for real and binding cash bets, and experiment 3 of Slovic in which gambles for values of up to $4 cash or nineteen packs of cigarettes were used.

THEORY 2: *Income Effects*. Three different modes of extracting subjects' attitudes have been used in existing experiments. Subjects were asked their preference between pairs[1] of lotteries, they were asked the maximum amount they would pay for the right to play various lotteries, and they were asked the minimum amount for which they would sell various lotteries. Clearly the income position of a subject can differ among these situations and this could account for apparent inconsistencies in preference. In three experiments income effects are of potential importance: Lichtenstein and Slovic (1971), experiment 3; Lichtenstein and Slovic (1973); and Slovic, experiment 3.

In Lichtenstein and Slovic (1971), experiment 3, subjects knew that all the gambles

[1]In Solvic subjects were asked to rank lotteries from sets of different sizes.

would be played at the end of the experiment. First, the subjects indicated their preferences from among a series of pairs of bets, the most preferred of which was to be played. After these choices the subjects were given a series of bets for which selling limit prices were extracted by standard techniques (see Gordon Becker, Morris DeGroot, and Jacob Marschak). After all choices were made, the relevant bets were played. Since all bets had a positive expected value, subjects had an increasing expected income throughout the experiment. Once one has agreed to play several *P* bets and expected income has accordingly increased, it is not so surprising that subsequently one is willing to go for riskier but potentially more rewarding gambles. Standard theories of risk taking suggest that risk aversion decreases with income, so as expected income increases, a tendency toward a higher limit selling price (the certainty equivalent for lotteries) would be expected. Thus the data which show preference reversals are consistent with an "income effect" hypothesis.

In Slovic, experiment 3, subjects first revealed indifference curves in a cigarette-money space. From this exercise they had an expectation of receiving some cigarettes (up to nineteen packs) from lotteries they knew would be played. A preference for a monetary dimension would thus be expected when two or three days later subjects were offered a choice between cigarette-money commodity bundles to which they had previously expressed indifference. Again the "income effect" hypothesis is consistent with the data.

The third case (Lichtenstein and Slovic, 1973) is an experiment conducted on a casino floor. Bets were played as soon as preferences were indicated. The wealth position of the subject at any time depended upon the sequence of wins and losses leading up to that time and these are not reported. Consequently, the relationship between the theory and this experiment cannot be determined.

THEORY 3: *Indifference*. In all experiments except those in Slovic, subjects were required to register a preference among bets. No indications of indifference were allowed. Thus the preference reversals exhibited in all other experiments could have been due to a systematic resolution of indifference on the part of subjects forced to record a preference. Slovic's results are also consistent with this hypothesis.

B. *Psychological-Theoretic Hypotheses*

THEORY 4: *Strategic Responses*. Everyone has engaged in trade and has some awareness of the various distributions of gains which accompany trade. Thus when asked to name a "selling price" an individual's natural strategic response may be to name a price above any true limit price modulated by what an opponent's preferences are conjectured to be. When asked to name a "buying price," one would tend to understate the values. This strategic reaction to the very words "selling price" may be very difficult to overcome even though the subject is selling to a passive lottery in which such strategic posturing may actually be harmful.

This theory predicts the reversals of all experiments except those reported in Slovic where the words buying and selling were not used. Notice that this theory would predict reversals when selling prices are elicited and fewer reversals, or reversals of the opposite sort, when buying prices are asked for. That is, one can argue that there is little ambiguity about the "value" of a *P* bet (for example, with probability of .99 win $4, and lose $1 with probability of .01). However, this is not true for the corresponding $ bets (for example, win $16 with probability one-third and lose $2 with probability two-thirds). Thus any tendency to state selling prices higher than the true reservation prices will primarily affect prices announced for $ bets. This behavior clearly can yield apparent preference reversals of the type reported. The same argument applied to buying prices suggests that there will be a tendency to understate the value of $ bets more than *P* bets.

Experiment 2 of Lichtenstein and Slovic (1971) used buying prices rather than selling prices. Compared with experiment 1 (which involved selling prices), Lichtenstein and Slovic report that for experiment 2 the rate of reversals was significantly lower (significance

level at least .01) and the rate of opposite reversals significantly higher (also at least .01). Further, they report that bids for the P bets average \$.07 below expected value in experiment 1, but \$.44 below expected value in experiment 2. Bids for the \$ bets were \$3.56 higher than expected value in experiment 1, and \$.04 below expected value in experiment 2. Thus, the data in these two experiments are quite consistent with this theory.

THEORY 5: *Probabilities*. With the exception of Slovic, experiments 1, 2, and 4, all experiments involved lotteries at some stage. Naturally if subjective probabilities are not well formed, or change during the experiment, consistency among choices is not to be expected. In fact, probabilities in all experiments except Lichtenstein and Slovic (1971), experiments 1 and 2, were operationally defined, and with the exception of Lichtenstein and Slovic (1973), there was no reason to suspect that they may have changed during the experiment.

THEORY 6: *Elimination by Aspect*. (See Amos Tversky, 1972.) Let A be a set of "aspects" and let the objects be subsets of A. This theory holds that individuals order the elements of A and then choose from among objects by a lexicographic application of this underlying order. Specifically, the stochastic version holds that an element x of A is chosen at random (perhaps with a probability proportional to its importance), and all objects B, such that $x \notin B$, are then eliminated. The process continued until only one object remains.

This theory runs counter to traditional economic reasoning on two counts. First the lexicographic application runs directly counter to the principle of substitution (quasi concavity of utility functions). Secondly, the random elimination choice process does not sit well with the idea of maximization or "rational" choice.

One implication of the model is a type of moderate stochastic transitivity.[2] The heart of the preference reversal phenomenon is shown above to be a type of cyclic choice. Such an intransitivity is in violation of the moderate stochastic transitivity property of the model. Thus the preference reversal phenomenon must be added to Tversky's own work (1969) as situations in which the elimination-by-aspects model does not seem to work.

THEORY 7: *Lexicographic Semiorder*. In a classic paper Tversky (1969) demonstrated that binary choices could cycle in a predictable fashion. The argument used was that choices are made on the basic dimensions of objects, but when for two objects the magnitudes along a dimension becomes "close," their magnitudes are treated as being equal. Thus a series of objects x, y, z, and w may be ordered as listed when compared in that order because each is close to those adjacent on a given dimension. Yet w may be chosen over x because the magnitudes on this dimension are far enough apart to be discernible.

It is difficult to see how this argument can be applied to account for the cycles in the reversal phenomenon. No long chains of binary comparisons were involved. No small magnitude differences, such as those used by Tversky, were present. We suspect that whatever ultimately accounts for the preference reversals will also account for the Tversky intransitivities, but we doubt that it will be the lexicographic semiorder model.

THEORY 8: *Information Processing—Decision Costs*. Individuals have preferences over an attribute space, but the location of an object in this attribute space may not be readily discernible. Resolution of choice problems, which involves locating an object in the attribute space, is a costly, disagreeable activity, so in their attempt to avoid decision costs people tend to adopt the following simple rule. An individual first looks at the "most prominent" dimension or aspect of the object. The magnitude of this aspect, called an "anchor," is used as the value of the object and is adjusted upward or downward to account for other features. As an empirical generalization the psychologists note that the adjustments are usually inadequate so the ultimate choice is heavily influenced by the starting point or

[2]If $P(x, y) \geq \frac{1}{2}$ and $P(y, z) \geq \frac{1}{2}$, then $P(x, z) \geq \min[P(x, y), P(y, z)]$.

anchor. Individuals who originally choose the *P* bet have tended to focus upon the probability of winning and inadequately adjust for the low monetary amounts. When asked about selling price or buying price, they naturally focus on dollars first and adjust for the probabilities. Since the adjustments for probabilities are inadequate, the dollar bets tend to be given the higher value. Thus, the "preference reversal" phenomenon is explained.

This theory is consistent with all experiments where no incentives were used. It is also consistent with the choices from among indifferent objects such as those in the Slovic 1975 study. When incentives are used, however, more effort in decision making is warranted and the frequency of reversals should go down. Thus on this theory one might have expected fewer reversals than occurred in the Lichtenstein and Slovic (1973) study, but since no control group (i.e., a group playing the gambles without monetary incentives) existed for this subject pool, the results are inconclusive.

THEORY 9: *Information Processing—Response Mode and Easy Justification.* The anchoring and adjustment mechanism described above may exist but it may be entirely unrelated to the underlying idea of decision-making costs. Indeed Lichtenstein and Slovic argue only that "variations in response mode cause fundamental changes in the way people process information, and thus alter the resulting decisions" (1971, p.16). The view is that of the decision maker "as one who is continually searching for systematic procedures that will produce quick and reasonably satisfactory decisions" (Slovic, p. 280). On occasion, it is argued that the mechanism is "easy to explain and justify to oneself and to others" (Slovic, p. 280). The anchoring process described above is offered as the mechanism that people adopt. The particular dimension used as an anchor is postulated to be a function of the context in which a decision is being made. Such thinking may not necessarily be contrary to preference theory. Rather, it is as though people have "true preferences" but what they *report* as a preference is dependent upon the terms in which the reporting takes place. Certain words or contexts naturally induce some dimensions as anchors while others induce other dimensions. The theory is consistent with all observations to date. Details can be found in Slovic.

C. *Experimental Methods*

The psychologists whose work we are reporting are careful scientists. Yet a bit of suspicion always accompanies a trip across a disciplinary boundary. In particular, we consider four possible sources of experimental bias.

THEORY 10: *Confusion and Misunderstanding.* In all experiments subjects were trained, rehearsed, and/or tested over procedures and options. In all instances repeated choices were made. In general there is reason to believe there was little confusion or misunderstanding, and in all cases the results hold up even when the responses of potentially confused subjects are removed from the data. However, there is some evidence reported in Lindman that suggests some type of "learning" takes place with experience. All experimenters reported some very "erratic" choices whereby, for example, a subject offered to pay more for a gamble than the maximum that a favorable outcome would yield.

THEORY 11: *Frequency Low.* If the phenomenon only occurs infrequently or with a very few subjects, there may not be a great need for concern or special attention. In fact, however, the behavior is systematic and the rate of reversals is high. Consider, for example, the following results, recalling that a *P* bet is a lottery with a high probability of winning a modest amount while the $ bet has a low probability of winning a relatively large amount of money. The Lichtenstein and Slovic (1971) study found that of 173 subjects indicating a preference for the *P* bet, 127 (73 percent) always placed a higher monetary valuation on the $ bet (they called these predicted reversals). On the other hand, the reverse almost never happens. That is, individuals who state that they prefer the $ bet will announce prices which are consistent with their choices. In this same study, for

example, 144 subjects *never* made the other reversal (termed unpredicted reversals).

THEORY 12: *Unsophisticated Subjects.* Psychologists tend to use psychology undergraduates who are required to participate in experiments for a grade. With the exception of Lichtenstein and Slovic (1973) the sources of subjects were not made explicit in the studies. If indeed psychology undergraduates were used, one would be hesitant to generalize from such very special populations.

THEORY 13: *The Experimenters were Psychologists.* In a very real sense this can be a problem. Subjects nearly always speculate about the purposes of experiments and psychologists have the reputation for deceiving subjects. It is also well known that subjects' choices are often influenced by what they perceive to be the purpose of the experiment. In order to give the results additional credibility, we felt that the experimental setting should be removed from psychology.

II. Experimental Design

Our format was designed to facilitate the maximum comparisons of results between experiments. The gambles used for our experiments (see Table 2) were the same ones used in Lichtenstein and Slovic (1971), experiment 3, where actual cash payoffs were made. They used a roulette wheel to play the gambles and, therefore, all probabilities were stated in thirty-sixths. The random device for our experiments was a bingo cage containing balls numbered 1–36. This eliminates the problem of nonoperational probabilities that was raised by Theory 5. All the gambles were of the form: if the number drawn is less than or equal to n, you lose \$$x$, and if the number drawn is greater than n, you win \$$y$.

The procedures for both of our experiments were so nearly identical that we shall describe only the first experiment in detail. Only those features of the second experiment that differ from the first will be discussed.

A. *Procedures: Experiment 1*

Student volunteers were recruited from economics and political science classes. They were told that it was an economics experiment, that they would receive a minimum of \$5, and that the experiment would take no longer than one hour. As the subjects arrived, they were randomly divided into two groups. The groups were in separate rooms, and there was no communication between them until after the experiment. Once the experiment was started, subjects were not allowed to communicate with each other though they were allowed to ask the experimenters questions.

Table 3 gives the organization of the experiment. At the start of the experiment the subjects received a set of general instructions that described the nature of the gambles they were to consider and explained how they were

TABLE 2—EXPERIMENT 1. PAIRS OF GAMBLES USED IN THE EXPERIMENTS

Pairs	Type	Probability of Winning	Amount if Win	Amount if Lose	Expected Value
1	*P*	35/36	\$ 4.00	−\$1.00	3.86
	\$	11/36	\$16.00	−\$1.50	3.85
2	*P*	29/36	\$ 2.00	−\$1.00	1.42
	\$	7/36	\$ 9.00	−\$ .50	1.35
3	*P*	34/36	\$ 3.00	−\$2.00	2.72
	\$	18/36	\$ 6.50	−\$1.00	2.75
4	*P*	32/36	\$ 4.00	−\$ .50	3.50
	\$	4/36	\$40.00	−\$1.00	3.56
5	*P*	34/36	\$ 2.50	−\$ .50	2.33
	\$	14/36	\$ 8.50	−\$1.50	2.39
6	*P*	33/36	\$ 2.00	−\$2.00	1.67
	\$	18/36	\$ 5.00	−\$1.50	1.75

TABLE 3—EXPERIMENT 1

Parts	Group 1 No Monetary Incentives	Group 2 Monetary Incentives
1	Preferences for Pairs (1), (2), (3)	
2	Selling Prices, All Twelve Gambles	
3	Preferences for Pairs (4), (5), (6)	

to be paid. These are included in the Appendix. Throughout the experiments all instructions and other materials handed out were read aloud. The instructions included a sample gamble (not used in the actual experiment): lose $1 if the number on the ball drawn is less than or equal to 12 and win $8 if the number is greater than 12. The way the gambles worked was demonstrated.

Two different monetary incentive systems were used which together control for Theory 1 and allow Theory 8 to be assessed. In one room (group 1) subjects were told that they would be asked to make a series of decisions concerning various gambles, and that when they were finished they would be paid $7. In the other room (group 2) subjects were told that at the end of the experiments one of their decisions would be chosen at random (using a bingo cage to determine which one) and their payment would depend upon which gamble they chose and upon the outcome of that particular gamble. It was explained that they had a credit of $7 and whatever they won or lost would be added to or substracted from that amount. Finally, it was stated that the most they could lose on any of the gambles was $2 so that $5 was the minimum possible payment.[3]

The use of a randomizing device to pick which decision "counted" was intended to reduce the problem of income effects discussed as Theory 2. Strictly speaking, even this procedure does not completely eliminate the possibility of some income effects, but it should reduce their magnitude substantially. Here there is little opportunity to have a growing expectation of rewards over the course of the experiment.

Part 1 of the experiment was distributed (the subjects were allowed to keep the instructions). This part consisted of three pairs of gambles.[4] For each pair subjects were told to indicate which bet they preferred or if they were indifferent. Subjects were told that if one of these three pairs was chosen at the end of the experiment, the two gambles would be played and that individual payments would be made according to which gamble was listed as preferred. Indifference was allowed and the subjects were told, "If you check 'Don't care,' the bet you play will be determined by a coin toss." Indifference was thus allowed and operationally defined in conformance with Theory 3.

After all subjects had completed part 1, the instructions and part 1 were collected and the instructions for part 2 were distributed. For part 2 of the experiments the subjects were asked to give their reservation prices for each of the twelve bets (the order of presentation was randomized). Specifically, subjects were asked "What is the *smallest* price for which you would sell a ticket to the following bet?"[5]

In order to ensure that actual reservation prices were revealed, the method suggested by Becker, DeGroot, and Marschak was employed. If one of the twelve items were chosen to be played at the end of the experiment, an offer price between $0.00 and $9.99 would be randomly generated and the subjects would play the gamble if their announced reservation price exceeded the offer price. Otherwise they would be paid the offer price (in addition to the $7 credit).[6] Thus our procedures

[3]This was the only difference in the instructions between the two rooms. In the other room also, a decision was to be chosen randomly at the end of the experiment. However, it was stated that this was just for fun as people often wish to know how much they would have won.

[4]In each pair the bets were referred to as *A* and *B* Assignment of *P* bets and $ bets to *A* or *B* was done randomly. On all materials passed out students were told to write their name, Social Security number, and in the room where payoffs were uncertain, their address.

[5]Announced preferences and those inferred from reservation prices should agree, but as this need not be the case with buying prices, no experiments involving buying prices were considered.

[6]The offer prices were generated by making three draws (with replacement) from a bingo cage containing balls numbered 0–9, these three draws giving the digits of the offer price.

conformed to those used in many other experiments and the problems raised by Theory 1 were avoided.

In order to be sure that all subjects fully understood how payments were to be determined, the instructions to part 2 were rather elaborate. The details, which can be found in the Appendix, include the following: an explanation about why it was in the subjects' best interest to reveal their true reservation prices; a practice gamble; a demonstration of the procedures; and a written test. The correct answers to the test were discussed and subjects' questions were answered. These procedures were designed to anticipate the problems raised by Theory 10. Subjects were allowed to work at part 2 at their own pace and were allowed to determine selling prices in whatever order they pleased.

Part 3 was identical to part 1 except that the remaining three pairs of bets were presented as shown on Table 3. Again, for each pair, subjects were asked to indicate a preference for bet A, bet B, or indifference. This procedure controls for a possible order effect implicit in the "cost of decision making" arguments of Theory 8. Once the subject has "invested" in a rule which yields a precise dollar value, then he/she would tend to use it repeatedly when the opportunity arises. Thus, we might expect greater consistency between decisions of parts 2 and 3 than between those of parts 1 and 2. After completing this part of the experiment, the subjects were paid as described.

B. *Procedures: Experiment 2*

The purpose of this experiment was to test the strategic behavior theory described as Theory 4. The structure of the experiment was identical to that of experiment 1 with two major exceptions. First, section 2 of the experiment was replaced at points by a section in which "limit prices" were extracted without references to market-type behavior. Second, subjects were not partitioned according to the method of payment. All subjects were paid with the same procedure as group 2 in experiment 1.

The organization of experiment 2 is shown in Table 4. Subjects were randomly divided into two rooms (the same two as used before) and designated as group 1 and group 2. Each group received identical instructions except the order in which the parts were administered was different as shown in Table 4.

TABLE 4—EXPERIMENT 2

Parts	Group 1	Group 2
	Monetary Incentives	
1	Preferences for Pairs (1), (2), (3)	
2	Selling Prices, All Twelve Gambles	Dollar Equivalents, All Twelve Gambles
3	Preferences for Pairs (4), (5), (6)	
4	Dollar Equivalents, All Twelve Gambles	Selling Prices, All Twelve Gambles

Part 2 for group 1 and part 4 for group 2 were identical to part 2 of experiment 1. Part 4 for group 1 and part 2 for group 2 consisted of a new section. In this new section no words suggestive of market-type activity (for example, selling prices and offer prices) were used. Instead students were asked to give "the exact dollar amount such that you are indifferent between the bet and the amount of money." For the operational details of how this was accomplished the Appendix should be consulted.

III. Results

A. *Experiment 1*

Table 5 summarizes the results for the room in which the subjects' payment was independent of their choices. It is clear that the reversal phenomenon has been replicated: of the 127 choices of P bets, 71 (56 percent) were inconsistent with the announced reservation prices. By comparison only 14 (11 percent) of the 130 choices of \$ bets were contradicted by the quoted reservation prices. Allowing the subjects to express indifference appears to have had little impact as in only 7 (3 percent) of the 264 choices made, was indifference indicated.

The propensity to reverse was the same for preferences obtained before and after selling prices for both types of bets. Thus, if asking

TABLE 5—FREQUENCIES OF REVERSALS, EXPERIMENT 1 (NO INCENTIVES)

	Bet	Choices	Reservation Prices		
			Consistent	Inconsistent	Equal
Total	*P*	127	49	71	7
	$	130	111	14	5
Indifferent		7			
Before Giving	*P*	73	30	39	4
Prices	$	56	48	5	3
After Giving	*P*	54	19	32	3
Prices	$	74	63	9	2
$n = 44$					

for selling prices focuses attention on the dollar dimension, it does not stay focused on it. The proportions in which *P* bets and $ bets were chosen before pricing differed significantly from those obtained after the prices (significant at .025 but not at .01). No other statistically significant effects were found.

Table 6 shows the corresponding data for the room in which the decisions were made for real money. Clearly (and unexpectedly) the preference reversal phenomenon is not only replicated, but is even stronger. Seventy percent of the choices of *P* bets were inconsistent with announced selling prices while reversals occurred for just 13 percent of the $ bet choices. Choice patterns and reversal rates appear to be the same for choices made before and after obtaining selling prices. The only significant differences between the performance in the two rooms are a higher proportion of selections of the $ bet in the incentive room (easily significant at .01 levels) and also a greater proportion of reversals on *P* bets (just clears the bar at .05).

We calculated a variety of summary statistics on the prices. The prices for $ bets tend to be higher than the prices for the corresponding *P* bets and were above their expected values. The distributions are apparently different for the two types of bets. In all twelve cases the mean, median, and estimated standard deviations were greater for the $ bet than for the corresponding *P* bet. There does not seem to be any systematic difference between the prices quoted in the two rooms. For each of the twelve bets the hypothesis of equal means was rejected only once (the *P* bet in pair number 2), and the t-statistic was just significant at a .05 level. From Table 7 one can see that not only were the preference reversals frequent, but also large. The magnitude of the reversals is generally greater for the predicted reversals than for the unpredicted reversals and also tends to be some-

TABLE 6—FREQUENCIES OF REVERSALS, EXPERIMENT 1 (WITH INCENTIVES)

	Bet	Choices	Reservation Prices		
			Consistent	Inconsistent	Equal
Total	*P*	99	26	69	4
	$	174	145	22	7
Indifferent		3			
Before Giving	*P*	49	15	31	3
Prices	$	87	70	12	5
After Giving	*P*	50	11	38	1
Prices	$	87	75	10	2
$n = 46$					

TABLE 7—EXPERIMENT 1: MEAN VALUES OF REVERSALS
(In Dollars)

	Predicted		Unpredicted	
Bet	Incentives	No Incentives	Incentives	No Incentives
1	1.71	2.49	.40	.79
2	1.45	2.64	.51	.90
3	1.48	1.29	1.00	.25
4	3.31	5.59	3.00	1.83
5	1.52	1.79	.38	1.29
6	.92	1.18	.33	.31

what smaller for the group with incentives "on." Thirty-four individuals (20 in the incentives room and 14 in the other) reversed every time they chose a P bet and of the 24 individuals who never reversed, 14 of them always chose the $ bet.

B. *Experiment 2*

Tables 8 and 9 summarize the results of experiment 2. Clearly the preference reversal phenomenon has again been replicated, and the strategic or bargaining behavior explanation shot down. If this explanation had been correct, reversals should have been obtained when using selling prices and not when dollar equivalents were asked for. It is apparent from the tables that this simply is not the case. Further, this theory would have predicted that selling prices should be higher than the monetary equivalents, but this is not true either. The mean selling price exceeded the mean equivalent in only ten of the twenty-four cases. Again, in every instance the mean price and dollar amount for a $ bet exceeds the respective means for the corresponding P bet. For each bet six t-tests(testing equality of means within and between rooms) were calculated. Of the seventy-two tests calculated the null hypothesis was rejected four times at a significance level of .05 and never at the .01 level. The overall conclusion is that the results obtained using prices and dollar equivalents are essentially the same. In both rooms and by both prices and equivalents approximately one-half the subjects reversed whenever they

TABLE 8—EXPERIMENT 2 SELLING PRICES
GROUP ONE

	Bet	Choices	Consistent	Inconsistent	Equal
				Selling Prices	
Total	P	44	8	30	6
	$	72	54	15	3
Indifferent		4			
Preferences	P	20	5	12	3
before Prices	$	39	24	12	3
Indifferent		1			
Preferences	P	24	3	18	3
after Prices	$	33	30	3	0
Indifferent		3			
				Equivalents	
Total	P	44	4	34	6
	$	72	59	11	2
Indifferent		4			
$n = 20$					

TABLE 9—EXPERIMENT 2: EQUIVALENTS
GROUP TWO

	Bet	Choices	Consistent	Inconsistent	Equal
				Equivalents	
Total	P	44	16	27	1
	$	64	54	9	1
Indifferent		0			
Preferences	P	22	8	14	0
before	$	32	27	4	1
Equivalents					
Preferences	P	22	8	13	1
after	$	32	27	5	0
Equivalents					
				Selling Prices	
Total	P	44	19	22	3
	$	64	51	10	3
$n = 18$					

chose a P bet. The number of individuals who chose a P bet at least once and never reversed varied between two and four.

IV. Conclusion

Needless to say the results we obtained were not those expected when we initiated this study. Our design controlled for all the economic-theoretic explanations of the phenomenon which we could find. The preference reversal phenomenon which is inconsistent with the traditional statement of preference theory remains. It is rather curious that this inconsistency between the theory and certain human choices should be discovered at a time when the theory is being successfully extended to explain choices of nonhumans (see John H. Kagel and Raymond C. Battalio, 1975, 1976).

As is clear from Table 1 our design not only controlled for the several possible economic explanations of the phenomena, but also for all but one of the psychological theories considered. Note that all the theories for which we exercised control can be rejected as explanations of the phenomena. Thus several psychological theories of human choice are also inconsistent with the observed preference reversals. Theory 8 is rejected since reversals do not go down as rewards go up. Theories 6 and 7 are rejected since the original results of Lichtenstein and Slovic (1971) have been replicated.

The one theory which we cannot reject, 9, is in many ways the least satisfactory of those considered since it allows individual choice to depend upon the context in which the choices are made. For example, if the mode of response or the wording of the question is a primary determinant of choice, then the way to modify accepted theory is not apparent. Even here, however, we have additional insight. If the questions give "cues" which trigger a mode of thinking, such cues do not linger. The reversals occur regardless of the order in which the questions are asked.

The fact that preference theory and related theories of optimization are subject to exception does not mean that they should be discarded. No alternative theory currently available appears to be capable of covering the same extremely broad range of phenomena. In a sense the exception is an important discovery, as it stands as an answer to those who would charge that preference theory is circular and/or without empirical content. It also stands as a challenge to theorists who may attempt to modify the theory to account for this exception without simultaneously making the theory vacuous.

APPENDIX

These instructions are those given to group 1 in experiment 2. From these, with the help of Tables 3 and 4 and the test, the instructions for all experiments can be reproduced. In

order to save space only those portions containing detailed instructions and examples used are shown. For example, part 1 consists of three similar items only one of which is shown.

Instructions

The experimenters are trying to determine how people make decisions. We have designed a simple choice experiment and we shall ask you to make one decision in each of several items. Each decision you shall make will involve one or more *bets*. If a bet is played, then one ball will be drawn from a bingo cage that contains 36 balls numbered 1, 2, . . . , 36. Depending upon the nature of the bet, the number drawn will determine whether you lose an amount of money or win an amount of money. Bets will be indicated by Figure 2. For example, if you play the following bet, then you will lose $1 if the number drawn is less than *or equal to* 12, and you will win $8 if the number drawn is greater than 12.

You will be paid in the following fashion. We first give you $7. After you have made a decision on each item, one item will be chosen at random by drawing a ball from a bingo cage. The bet(s) in the chosen item will then be played. You will be paid an amount depending upon your decisions and upon the outcomes of the bets in the chosen item—any amount you win will be added to the $7, and any amount you lose will be subtracted from the $7. However, the most you can lose on a bet is $2, so you will certainly receive at least $5. All actual payments will occur after the experiment.

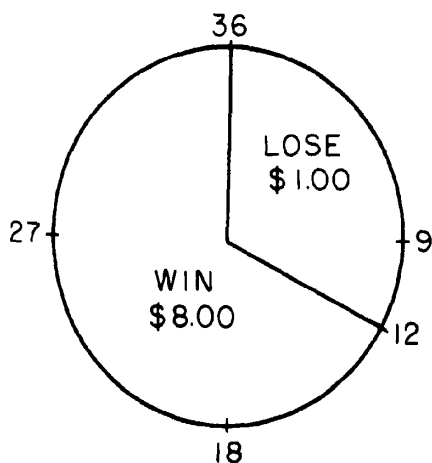
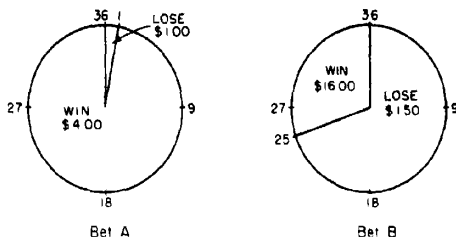



FIGURE 2

FIGURE 3

PART 1: If an item from this part is chosen at the end of the experiment, you will play the bet you select. If you check "Don't care," the bet you play will be determined by a coin toss.

Item 1: Consider carefully the following two bets shown in Figure 3.

Suppose you have the opportunity to play one of these bets. Make *one* check below to indicate which bet you would prefer to play:

Bet *A*:_______

Bet *B*:_______

Don't care:____

. . . the instructions continue to items 2 and 3 from Table 2 . . .

PART 2: *Instructions*: In each of the items below, you have been presented a ticket that allows you to play a bet. You will then be asked for the *smallest* price at which you would sell the ticket to the bet.

If an item from this part is chosen at the end of the experiment, we will do the following. First, a bingo cage will be filled with 10 balls numbered 0, 1, 2, . . . , 9. Then 3 balls will be drawn from this cage, with each ball being replaced before the next is drawn. The numbers on these 3 balls will determine the digits of an offer price between $0.00 and $9.99, with the first number being the penny (right) digit, the second number the dime (middle) digit, and the third number the dollar (left) digit. If this offer price is greater than or equal to the price you state is your

minimum selling price for the item's bet, you would receive the offer price. If the offer price is less than your selling price, you would play the bet and be paid according to its outcome.

It is in your best interest to be accurate; that is, the best thing you can do is to be honest. If the price you state is too high or too low, then you are passing up opportunities that you prefer. For example, suppose you would be willing to sell the bet for $4 but instead you say that the lowest price you will sell it for is $6. If the offer price drawn at random is between the two (for example $5) you would be forced to play the bet even though you would rather have sold it for $5. Suppose that you would sell it for $4 but not for less and that you state that you would sell it for $2. If the offer price drawn at random is between the two (for example $3) you would be forced to sell the bet even though at that price you would prefer to play it.

Practice Item: What is the *smallest* price for which you would sell a ticket to the following bet? _____. (The group is then shown the same bet as Figure 2.)

Item 0:[7] What is the *smallest* price for which you would sell the following bet shown in Figure 4?_____

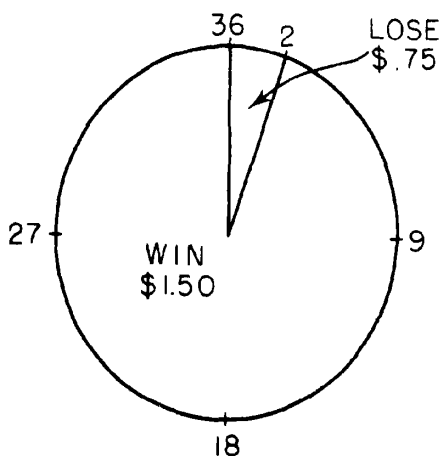



FIGURE 4

[7]Item 0, listed here, and item 00, with $7 with 15/36, lose $1.25 with 21/36, were given as a "test" prior to undertaking part 2.

The offer price is $_. _ _.
The number drawn for the bet was _____.

If this item had actually been played, the amount I would (circle the correct word) gain lose is _____.
... (see fn. 7) ...

Item 4: What is the *smallest* price for which you would sell a ticket to the following bet shown in Figure 5? _____

... continue with all items in Table 2 ...

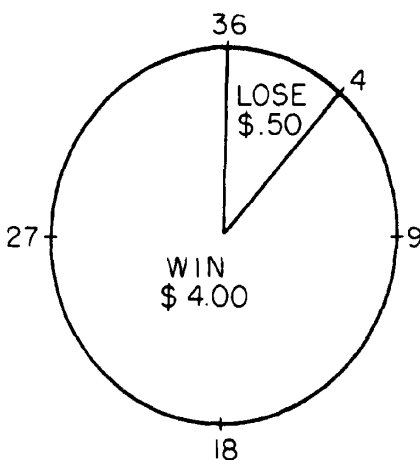



FIGURE 5

PART 3: The items below are like the items of part 1. If one of them is chosen at the end of the experiment, you will play the bet you select. If you check "Don't care," then the bet you play will be determined by a coin toss.

Item 16: Consider carefully the following two bets shown in Figure 6:

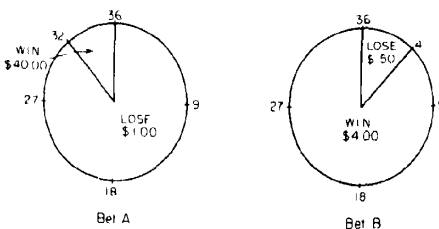



FIGURE 6

Suppose you have the opportunity to play one of these bets. Make *one* check below to indicate which bet you would prefer to play:

Bet *A*:________

Bet *B*:________

Don't care:____

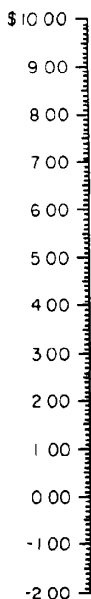
...continue with items 4, 5, and 6, Table 2...

PART 4: *Instructions*: Each of the items in this part shows a bet and a monetary scale. As in the example below the dollar amounts increase as you go from the bottom to the top of the scale. (The group is then shown the same bet in Figure 2 plus the scale shown in Figure 7.)

For each item in this part we ask you to do the following. Put your finger at the bottom of the scale and ask yourself which you would prefer to have—the bet shown or the dollar amount. In this case the bet offers 24 chances out of 36 of winning $8 and 12 chances of losing $1. We assume you prefer the bet to giving up $2. Now move your finger up the scale towards the top continuing to ask the same question. At the very top of the scale is an amount of money greater than that which could be won on the bet. We assume you would prefer the $10 to the bet. All scales in this part will be constructed so that for some of the numbers at the bottom you will prefer to have the bets and for some at the top you will prefer to have the money. What we would like to know is this: what is the exact dollar amount such that you are indifferent between the bet and the amount of money. Mark this amount (with an X) on the scale. Since X's are not always easy to read, and as the scale may not be fine enough for you, we also ask that you write the amount checked in the space provided.

FIGURE 7

In order to provide you with an incentive to be as accurate as possible, we will do the following. If an item from this part is chosen, we will randomly choose one of the numbers shown on the scale. For example, for this scale a bingo cage would be filled with 10 balls numbered 0, 1, 2, . . . , 9. Then 3 balls would be drawn with replacement. The numbers on these 3 balls will determine an amount of money with the first ball drawn being the penny (right) digit, the second number the dime (middle) digit, and the third number the dollar (left) digit. If this number is greater than the amount you check, you will receive the number *drawn*. If the number is less than the amount checked, we will play the bet and you will be paid according to its outcome. If the number drawn is the same as the amount checked, the toss of a fair coin will determine whether you play the bet or get the money. As in this example, we will never generate any negative numbers, but all positive numbers shown on the scale will be equally likely.

Notice that your best interest is served by accurately representing your preference. The best thing you can do is be honest. If the number you mark is too high or too low, then you are passing up opportunities that you prefer. For example, suppose $4 is your point of indifference but you marked $6. If the amount of money drawn at random is anything between the two (for example, $5), you would be forced to play the bet even though you would rather have the drawn amount. Suppose your point of indifference was $4 and you marked $2. If the amount of money drawn at random is between the two

(for example, $3) then you would be forced to take the money even though you prefer to play the bet.

Item 0: (The group is shown the bet in Figure 4 and the monetary scale in Figure 7.)

The dollar amount drawn was $__.__ The number drawn for the bet was _____.

If this item had actually been played, the amount I would (circle the correct word) gain lose is ______.

. . . see fn 7 . . .

PART 5:

Item 19: On the scale mark the exact dollar amount such that you are indifferent between the bet and the amount of money. (The group is shown the same bet as Figure 5, and the monetary scale in Figure 7.) . . . continue with Items 1 through 6, Table 2 . . . .

REFERENCES

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *76*, 169–217.

______, "A Theory of Marriage: Part I," *J. Polit. Econ.*, July/Aug. 1973, *81*, 813–46; "Part II," Mar./Apr. 1974, suppl., *82*, 511–26.

G. M. Becker, M. H. DeGroot, and J. Marshak, "Measuring Utility by a Single-Response Sequential Method," *Behav. Sci.*, July 1964, *9*, 226–32.

I. Ehrlich, "The Deterrent Effect of Capital Punishment: A Question of Life and Death," *Amer. Econ. Rev.*, June 1975, *65*, 397–417.

R. C. Fair, "A Theory of Extramarital Affairs," *J. Polit. Econ.*, Feb. 1978, *86*, 45–61.

D. Hammermesh and N. Soss, "An Economic Theory of Suicide," *J. Polit. Econ.*, Jan./Feb. 1974, *82*, 83–98.

J. H. Kagel and R. C. Battalio, "Experimental Studies of Consumer Demand Behavior Using Laboratory Animals," *Econ. Inquiry*, Mar. 1975, *13*, 22–38.

______ **and** ______, "Demand Curves for Animal Consumers," mimeo., Washington Univ., St. Louis 1976.

S. Lichtenstein and P. Slovic, "Reversal of Preferences Between Bids and Choices in Gambling Decisions," *J. Exper. Psychol.*, July 1971, *89*, 46–55.

______ **and** ______, "Response-Induced Reversals of Preferences in Gambling: An Extended Replication in Las Vegas," *J. Exper. Psychol.*, Nov. 1973, *101*, 16–20.

H. R. Lindman, "Inconsistent Preferences among Gambles," *J. Exper. Psychol.*, Aug. 1971, *89*, 390–97.

P. Slovic, "Choice Between Equally Valued Alternatives," *J. Exper. Psychol.: Hum. Percep. and Perform.*, Aug. 1975, *1*, 280–87.

A. Tversky, "Intransitivity of Preferences," *Psychol. Rev.*, Jan. 1969, *76*, 31–48.

______, "Elimination by Aspects: A Theory of Choice," *Psychol. Rev.*, July 1972, *79*, 281–99.

Money, Bonds, and Foreign Exchange

By EUGENE F. FAMA AND ANDRÉ FARBER*

In the theory of finance, capital markets provide securities which allow the ownership of productive assets to be allocated among investors according to their willingness to bear risk in exchange for return, and independently of their tastes for the productive services provided by the assets. This paper uses the theory of finance to study the role of money as a productive asset and the transformation of money into a portfolio asset. True to its purpose, the capital market provides securities—nominal bonds—that allow an allocation of the current services of the money supply according to its current productivity and without regard for its future purchasing power risk, and an allocation of the future purchasing power risk of the money supply according to the willingness of investors to bear this risk in exchange for return and without regard for investor needs for the future transactions services of money. In open international capital markets, nominal bonds allow the purchasing power risk of a country's money supply to be borne by the residents of other countries, even though the money might provide transactions services only within the borders of the country issuing it.

Studying the links between money and nominal bonds sheds light on various issues in monetary theory and international finance. For example, it leads to a theory of money demand that differs from other approaches. It also allows us to show how differential purchasing power risks of the money supplies of different countries can lead to premiums in forward exchange rates as predictors of future spot exchange rates. Such premiums can exist even in situations where exchange rate uncertainty is irrelevant in the sense that it has no effect on how an investor splits his portfolio funds between local and foreign securities.

I. The Economic Environment

Initially we treat a one-country, two-good world, where the two goods are money and a commodity. The role of money is to reduce costs in exchanges of the commodity among individuals and firms. The world is multiperiod and all decisions are taken at discrete points in time. At any such time t, individuals (households) combine money and the commodity to produce consumption. Although the role of money is to reduce transactions costs, it is analytically convenient to avoid explicit consideration of such costs. Instead we summarize the properties of money as a medium of exchange in terms of a household production function for consumption, $c_t(m_t, q_t)$, which is assumed to be strictly increasing and strictly concave in m_t and q_t, the quantities of real money and of the commodity used by the household to produce consumption at time t.[1]

Real money m_t is

$$m_t = M_t \pi_t \tag{1}$$

where M_t is units of nominal money (for example, dollars) and π_t is the purchasing power of a unit of money, the number of units of the commodity that can be purchased with a unit of money. To avoid explicit consideration of how the transactions process leads to some average holding of money, we assume that money used to produce consumption does not evaporate in the process; it reappears as part of the individual's nominal wealth at t + 1. The units of the commodity q_t allocated to consumption at t do disappear. In fact, for simplicity the commodity is also taken to be

*Graduate School of Business, University of Chicago, and Université Libre de Bruxelles, respectively. This research is supported by a grant from the National Science Foundation. We have profited from the comments of G. Borts, A. Drazen, N. Gonedes, R. Ibbotson, R. Kormendi, S. Schaefer, M. Scholes, G. W. Schwert, and a referee. Special thanks are due to J. Frenkel and M. H. Miller.

[1]Stanley Fischer (1974) discusses the rationale for modelling money as an input in a production function and the equivalence of this approach to including money as an argument in a utility function.

the consumption good. The interpretation of the production function $c_t(m_t, q_t)$ is that, because of frictions in the transactions process, units of the commodity used in consumption need not show up one-for-one as units of consumption. Money is, in effect, a nonevaporating grease that helps to reduce the losses from friction.

There is a similar story about the role of money in the production decisions of firms. Firms combine money and the commodity at time t in production decisions that yield uncertain quantities of the commodity at future points in time. This uncertainty about the future outputs generated by current production decisions leads to uncertainty in the returns on common stock. For simplicity, our analysis of money concentrates on the decisions of individuals.

Non-interest-bearing fiat money is produced by the government at zero cost and sold to individuals and firms. The government uses the resources, units of the commodity, that it gets from the sale of money to help sustain the general economic environment (police, defense, etc.). The government does not levy taxes, nor does it make any transfer payments to individuals and firms.

Money has real value because of frictions in the system which are reduced when money is used as the medium of exchange. We assume that all such frictions are associated with the commodity, that is, with the production of consumption by households and with the production of the commodity by firms. In contrast, since we use the standard financial models of capital market equilibrium, we must assume that the capital market is frictionless; that is, securities are infinitely divisible and traded costlessly among investors. A more palatable assumption might be that transactions costs in carrying out barter exchanges of financial assets are trivial relative to transactions costs in barter exchanges of commodities, but such an assumption is cumbersome to model explicitly. One can say that it appears in an approximate form in our model through the assumption that costless exchanges of money and securities can only take place at discrete points in time.

At time t the purchasing power of money at the next point in time, t + 1, is uncertain; that is, $\tilde{\pi}_{t+1}$ is a random variable, denoted by the tilde. The perceived future needs of the government to sustain the economic environment, the major determinant of future supplies of money, are uncertain at t. Likewise, future demands for money services will depend on variables, like consumer-investor wealth, whose future values are uncertain at t. Much of what is interesting about the links between the markets for money and bonds has to do with the pricing of the uncertainty in the future value of money.

Having described the economic environment, the next step is a general analysis of the consumption-portfolio problem facing an individual at some time t and the role of money and nominal bonds in this decision. We then consider the pricing of nominal bonds when equilibrium in the capital market is as described by the models of William Sharpe and John Lintner. Finally, the analysis is extended to the pricing of money-related instruments in international money and capital markets.

II. The Role of Money in Consumption-Portfolio Decisions

The decision problem facing an individual is assumed to be as follows. At any time t, the individual must allocate his real wealth w_t to the production of consumption for period t and to investment in some portfolio of securities. The real value of his portfolio at t + 1 is his wealth at t + 1 which again must be allocated to consumption and investment.

To simplify the story, assume that there are only two types of securities available at time t; common stock, issued by firms, and one-period nominal bonds, which are free of default risk and can be issued by either individuals or firms. For current purposes no harm is done if we also assume that there is just a single issue of common stock available. The budget constraint facing the individual at time t is then

$$w_t = m_t + q_t + b_t + z_t \tag{2}$$

where b_t is real wealth allocated to nominal bonds, z_t is real wealth allocated to common

stock, and m_t and q_t are the quantities of real money and the commodity allocated to the production of consumption.

Recall that the units of the commodity q_t allocated to consumption at t are used up, whereas money allocated to the production of consumption reappears as part of wealth at t + 1. Thus, if R_{t+1} is the risk-free nominal rate of interest set at t for payment at t + 1, $\tilde{s}_{t+1}$ is the uncertain real rate of return on common stock from t to t + 1, and $\tilde{\delta}_{t+1} = (\tilde{\pi}_{t+1}/\pi_t) - 1$ is the uncertain rate of change in the purchasing power of money, then the real wealth at time t + 1 generated by the individual's consumption-portfolio decision at t is

$$(3) \quad \tilde{w}_{t+1} = m_t(1 + \tilde{\delta}_{t+1}) + b_t(1 + R_{t+1})(1 + \tilde{\delta}_{t+1}) + z_t(1 + \tilde{s}_{t+1})$$

where

$$(4) \quad m_t(1 + \tilde{\delta}_{t+1}) = (m_t/\pi_t)\tilde{\pi}_{t+1} = M_t\tilde{\pi}_{t+1}$$

is the real value at t + 1 of money carried forward from t, and $(1 + R_{t+1})(1 + \tilde{\delta}_{t+1})$ is the real value at t + 1 of a unit of real wealth allocated to nominal bonds at t. Like the real value of money at t + 1, the real return on nominal bonds at t + 1 is uncertain at t because of the uncertainty about the rate of change in the purchasing power of money.

At any time t the individual's tastes are assumed to be characterized by a von Neumann-Morgenstern utility function $U(c_t, w_{t+1})$. Following the analysis of Fama, $U(c_t, w_{t+1})$ is meant to be a derived utility function which shows the maximum expected utility of lifetime consumption, given that current consumption is c_t, next period's wealth is w_{t+1}, and optimal consumption-investment decisions are made at t + 1 and at all future points in time. Moreover, in Fama's analysis, if the individual's utility function for lifetime consumption is strictly increasing and strictly concave in its arguments, that is, if the individual is risk averse, then the derived function $U(c_t, w_{t+1})$ is strictly increasing and strictly concave in its arguments, so that the risk-aversion property is conserved.

To get the formal solution to the individual's decision problem at time t, we form the Lagrangian

$$(5) \quad L = \int_{\delta_{t+1}s_{t+1}} U(c_t(m_t, q_t), m_t(1 + \delta_{t+1}) + b_t(1 + R_{t+1})(1 + \delta_{t+1}) + z_t(1 + s_{t+1}))dF(\delta_{t+1}, s_{t+1}) + \lambda(w_t - m_t - q_t - b_t - z_t)$$

where λ is the Lagrange multiplier for the wealth constraint of (2) and $F(\delta_{t+1}, s_{t+1})$ is the joint distribution function for the random variables $\tilde{\delta}_{t+1}$ and $\tilde{s}_{t+1}$. Differentiating (5) with respect to each of the decision variables and setting the derivatives equal to zero, we obtain the general optimality condition that the marginal expected utility of a unit of wealth allocated to any of its four possible uses at t is the same for all possible uses:[2]

$$(6) \quad \frac{\partial L}{\partial m_t} = \frac{\partial c_t}{\partial m_t}\int \frac{\partial U}{\partial c_t} dF(\delta_{t+1}, s_{t+1}) + \int \frac{\partial U}{\partial w_{t+1}}(1 + \delta_{t+1})dF(\delta_{t+1}, s_{t+1}) - \lambda = 0$$

$$(7) \quad \frac{\partial L}{\partial q_t} = \frac{\partial c_t}{\partial q_t}\int \frac{\partial U}{\partial c_t} dF(\delta_{t+1}, s_{t+1}) - \lambda = 0$$

$$(8) \quad \frac{\partial L}{\partial b_t} = (1 + R_{t+1})\int \frac{\partial U}{\partial w_{t+1}}(1 + \delta_{t+1}) \cdot dF(\delta_{t+1}, s_{t+1}) - \lambda = 0$$

$$(9) \quad \frac{\partial L}{\partial z_t} = \int \frac{\partial U}{\partial w_{t+1}}(1 + s_{t+1}) \cdot dF(\delta_{t+1}, s_{t+1}) - \lambda = 0$$

There are many ways that these equations can be manipulated to gain insights into the details of an optimal consumption-portfolio decision. However, our specific interest is to study the links between the individual's decision with respect to bonds and the way money and the commodity are optimally combined to produce current consumption. Note first that,

[2]The fact that the functions $U(c_t, w_{t+1})$ and $c_t(m_t, q_t)$ are strictly increasing and strictly concave in their arguments implies that the first-order conditions of (6)–(9) lead to a maximum of expected utility. All functions are assumed to be differentiable to whatever extent is required by the analysis. In addition, we assume that the production function $c_t(m_t, q_t)$ and the utility function $U(c_t, w_{t+1})$ are such that negative values of c_t, m_t, and q_t are never optimal.

except for the term $(1 + R_{t+1})$, the expression for the marginal expected utility of b_t in (8) is the same as the second term in the expression for the marginal expected utility of m_t in (6). Except for the foregone interest, a unit of wealth carried forward from t to t + 1 in the form of money is the same as a unit of wealth carried forward in the form of a nominal bond. It follows that an optimal consumption-portfolio decision at t involves choosing real money m_t and bonds b_t so that the marginal expected utility of the marginal consumption produced by money at t is equal to the marginal expected utility associated with the interest earned at t + 1 on the wealth allocated to nominal bonds at t. Formally, equating (6) and (8) we find that

$$(10)\quad \frac{\partial c_t}{\partial m_t}\int \frac{\partial U}{\partial c_t}\, dF(\delta_{t+1}, s_{t+1}) = R_{t+1}\int \frac{\partial U}{\partial w_{t+1}}(1 + \delta_{t+1})\, dF(\delta_{t+1}, s_{t+1})$$

Equation (10) can also lead to insights into how money m_t and the commodity q_t should be combined to produce consumption at t. Using (10) to replace the second term in (6) and then equating the resulting expression with (7), we get

$$(11)\quad \frac{\partial c_t/\partial m_t}{\partial c_t/\partial q_t} = \frac{R_{t+1}}{1 + R_{t+1}}$$

The appearance of the nominal interest rate in (11) has an intuitive interpretation. In terms of risk, the directly comparable alternative to carrying money forward from t to t + 1 is to invest in nominal bonds. Thus, nominal bonds provide the natural reference point for the $c_t(m_t, q_t)$ decision. The opportunity loss in allocating a unit of real wealth to the commodity q_t instead of to nominal bonds is $(1 + R_{t+1})(1 + \tilde{\delta}_{t+1})$, the foregone real value of the wealth that would be realized from the bonds at t + 1. The opportunity loss in allocating a unit of real wealth to money instead of to nominal bonds is just $R_{t+1}(1 + \tilde{\delta}_{t+1})$ since the money shows up as part of wealth at t + 1. Equation (11) then says that the optimal way to combine m_t and q_t to produce a given level of consumption at t involves balancing money and the commodity so that the ratio of their marginal productivities is the ratio of their opportunity costs.

III. The Role of Nominal Bonds

This intuitive interpretation of (11) does not explain its important implication that the way m_t and q_t are optimally combined to produce consumption at t does not depend on the individual's tastes, in particular, his attitudes toward risk. This result seems counterintuitive since money used to produce consumption at t shows up as part of the individual's wealth at t + 1, but with uncertain purchasing power. It would seem that the individual's decision to use money for consumption at t forces him to bear the purchasing power risk of the money at t + 1 and that this should influence his money decision.

This conclusion overlooks the fact, captured in the mathematics of equations (6)–(11), that the nominal bond market can be used to separate the decision about how much of the purchasing power risk of money the individual will bear from the decision about how much money to hold for the production of consumption. A nominal bond is a forward contract on money, a promise to deliver a fixed amount of money at the next point in time in exchange for money now. If an individual allocates m_t of real wealth at t to money for the production of consumption during period t, the market realizes that the individual has $M_t = m_t/\pi_t$ in money which he can deliver for certain at t + 1. A perfect capital market would be willing to pay $M_t/(1 + R_{t+1})$ in money at time t, equivalent to $m_t/(1 + R_{t+1})$ units of the commodity, for the right to the money that the individual must carry forward to t + 1 as a consequence of his consumption decision. When the individual allocates $m_t = M_t\pi_t$ of real money to the production of consumption at t, this only reduces the real wealth that he can allocate to portfolio investments by $M_tR_{t+1}/(1 + R_{t+1})$, so that $M_tR_{t+1}/(1 + R_{t+1})$ is the cost of the current services of money.

Analogous statements apply to the money decisions of firms. Thus, as in any case where

there is a forward market for a durable asset, the simultaneous purchase of money in the spot market and forward sale of money in the nominal bond market allows one to rent money, that is, purchase its current transactions services, without bearing the uncertainty of its future purchasing power. In this way nominal bonds allow an optimal allocation of the current transactions services of the money supply and a separate optimal allocation of the money supply as a store of uncertain future value. There is a transactions demand for the current services of money, and then the nominal bond market provides the mechanism whereby the future services of money are capitalized and held as an interest-earning portfolio asset. Through the nominal bond market the purchasing power risk of the money supply can be allocated across investors in accordance with their willingness to bear risk in exchange for a return, and independently of their demands for current and future money services.[3]

This story about money and nominal bonds certainly depends to some extent on the assumption that there are discrete points in time when switches between money and bonds are costless. For example, if there are costs in switching between money and bonds, then the extent to which one issues nominal bonds to hedge uncertainty about the future purchasing power of current money holdings will no longer be completely unaffected by current estimates of future demands for money services. On the other hand, the oversimplifications in our model also cause us to understate the possibilities for hedging money holdings that are provided by the nominal bond market. For example, since we only consider one-period nominal bonds, our analysis overlooks the rich range of possibilities for hedging multiperiod money holdings that are provided by multiperiod bonds.

Finally, the role of nominal bonds in reallocating the purchasing power risk of the money supply is not well appreciated in the literature. For example, Fischer (1975) treats a model in which portfolio assets include nominal bonds and risk-free real bonds. He emphasizes that nominal bonds are dominated by real or indexed bonds when all investors are risk averse and agree on the stochastic properties of the economic environment. However, this result stems from the fact that Fischer deals primarily with the case where all bonds, nominal and real, arise solely from borrowing and lending among individuals so that their net stocks are zero. Thus, he overlooks the fact that the money supply is in effect a net stock of nominal bonds with nominal payoff next period equal to the current face value of the money supply.

Other studies of indexed bonds, for example, that of Nissan Liviatan and David Levhari, are aware that borrowing in the nominal bond market can be used to hedge holdings of money. However, since these studies are primarily concerned with the role of indexed bonds, the links between money and nominal bonds are never developed in detail. Our work attempts to provide some of the relevant details.

IV. Restatement in Terms of the Sharpe-Lintner Model

We now examine the results of the preceding section in the context of the model of capital market equilibrium of Sharpe and Lintner. This model gives us a specific example of how the pricing of nominal bonds induces investors to bear the purchasing power risk of the money supply. Moreover, the model is especially useful later when we apply our analysis to the determination of forward exchange rates in international money markets.

A. *Allocation of the Purchasing Power Risk of the Money Supply*

In our version of the Sharpe-Lintner model, portfolio assets include one-period risk-free real bonds as well as one-period nominal bonds and common stock. A special feature of the model is that optimal portfolios, the port-

[3]In contrast, in models such as those of Milton Friedman or James Tobin, there is a direct demand for money as a portfolio asset, and the current money decision is affected by comparisons of the risks and returns from holding money with the risks and returns on other assets.

folios relevant for choice by investors, involve combining an investment in the market portfolio of total invested wealth with borrowing or lending at the risk-free real rate of interest. Our interest is in showing how the money supply gets included in the market portfolio in such a way that the allocation of the purchasing power risk of the money supply among investors is independent of the allocation of its current transactions services.

As in the earlier analysis, money allocated to the production of consumption at time t is not used up in the process, and the money can be sold forward in the nominal bond market to increase portfolio investments at time t. The individual's investment in the market portfolio is his total wealth at t, less resources allocated to risk-free real lending (or plus resources received from borrowing at the risk-free real rate), less resources allocated to the commodity and to money for producing consumption during period t, plus the value of this money when sold forward in the nominal bond market for delivery at t + 1. The individual allocates this total to portfolio assets so as to replicate their weights in the market portfolio of total invested wealth.

Total invested wealth at any time t includes the total value of all the securities of all firms. However, it also includes the bond equivalent value of the household money supply, that is, the value of the money supply used by individuals to produce consumption, sold forward in the nominal bond market. The bond equivalent value of the part of the money supply held by firms shows up in the market value of their securities at t, whether or not firms actually issue nominal bonds against their holdings of money.

Since each investor holds the market portfolio (but in combination with different amounts of risk-free real borrowing or lending), each investor ends up bearing all the uncertainties associated with portfolio assets, including the purchasing power risk of the total money supply, in the proportions in which they are outstanding in the market. In other words, the bond market achieves an allocation of the money supply as a portfolio asset which is independent of the allocation of the current services of money as an input to the production decisions of households and firms.

B. *Equilibrium Real Bond Prices*

In the Sharpe-Lintner model, the real price of any portfolio asset j at time t is

$$(12)\quad P_{jt} = \frac{E(\tilde{P}_{j,t+1}) - \phi\, cov\,(\tilde{P}_{j,t+1}, \tilde{P}_{A,t+1})}{1 + r_{F,t+1}}$$

where $r_{F,t+1}$ is the risk-free real rate of interest for the period from t to t + 1, $\tilde{P}_{j,t+1}$ is the uncertain real value of asset j at t + 1, $\tilde{P}_{A,t+1}$ is the aggregate real value at t + 1 of the economy's invested wealth carried forward from t (including the real value at t + 1 of the money supply outstanding at t), $E(\tilde{P}_{j,t+1})$ is the expected value of $\tilde{P}_{j,t+1}$, $cov\,(\tilde{P}_{j,t+1}, \tilde{P}_{A,t+1})$ is the covariance between $\tilde{P}_{j,t+1}$ and $\tilde{P}_{A,t+1}$, and ϕ is the market price of risk. The parameter $cov\,(\tilde{P}_{j,t+1}, \tilde{P}_{A,t+1})$ is the market risk of asset j in the sense that it is the contribution of a unit of the asset to the variance of the total value of invested wealth; that is,

$$(13)\quad \sigma^2(\tilde{P}_{A,t+1}) = \sum_j \sum_k n_j n_k \cdot cov\,(\tilde{P}_{j,t+1}, \tilde{P}_{k,t+1}) = \sum_j n_j\, cov\,(\tilde{P}_{j,t+1}, \tilde{P}_{A,t+1})$$

where n_j is the number of units of asset j outstanding at t. All of the variables in (12) and (13) are measured in units of the commodity. Although money is the medium of exchange, the goal of investment is eventual real consumption, so it is appropriate to state the market equilibrium condition in real terms.

We can use equation (12) to determine the real price at time t of a one-period risk-free nominal bond. For simplicity, suppose the contract is for delivery of one unit of money at t + 1. Applying (12), we have

$$(14)\quad P_t = \frac{E(\tilde{\pi}_{t+1}) - \phi\, cov\,(\tilde{\pi}_{t+1}, \tilde{P}_{A,t+1})}{1 + r_{F,t+1}}$$

It is well to recall that the real value at t + 1 of the money supply carried forward from t is

part of $\tilde{P}_{A,t+1}$. As for any other component of real wealth at t + 1, $cov(\tilde{\pi}_{t+1}, \tilde{P}_{A,t+1})$ measures the contribution of the uncertainty about the real value of a unit of money carried forward to t + 1 to $\sigma^2(\tilde{P}_{A,t+1})$, the risk of total invested real wealth at t + 1. Since the money supply is part of invested wealth, the purchasing power risk of the money supply must be borne, and this is accomplished through the incorporation of a risk adjustment in the price of a nominal bond.

Another way to make this point is to compare the expected real rate of return on a nominal bond, call it $E(\tilde{r}_{t+1})$, with the risk-free real rate of interest $r_{F,t+1}$. The expected real rate of return on a nominal bond that pays one unit of money at t + 1 is

$$(15) \qquad 1 + E(\tilde{r}_{t+1}) = \frac{E(\tilde{\pi}_{t+1})}{P_t}$$

or, substituting (14) for P_t,

$$(16) \quad 1 + E(\tilde{r}_{t+1}) = \frac{1 + r_{F,t+1}}{1 - \phi\, cov(\tilde{\pi}_{t+1}, \tilde{P}_{A,t+1})/E(\tilde{\pi}_{t+1})}$$

We can see that there is an adjustment for risk in the expected real return on a nominal bond, that is, $E(\tilde{r}_{t+1}) \neq r_{F,t+1}$, and the difference between $E(\tilde{r}_{t+1})$ and $r_{F,t+1}$ depends on the purchasing power risk of the nominal bond, $cov(\tilde{\pi}_{t+1}, \tilde{P}_{A,t+1})$, per unit of its expected future purchasing power, $E(\tilde{\pi}_{t+1})$.[4]

V. Money, Bonds, and International Capital Markets

The analysis of the links between money and nominal bonds can be used to shed light on issues in international finance. We examine an open multicountry world in which there are no restrictions on the trading of securities and the commodity among countries. Each country is similar to the one we treat above in the sense that a single commodity is used by individuals as an input for producing consumption and by firms as an input to the production of uncertain future quantities of the commodity. The commodity is homogeneous across countries. The capital market of each country includes common stocks, one-period risk-free real bonds and one-period nominal bonds whose money payoffs are certain. Each country has its own money which is used as medium of exchange, and which we again model by making money an input to production by households and firms.

The money of a given country is used only by the residents (individuals and firms) of that country. This does not mean that the purchasing power risk of a country's money supply is borne entirely by its residents. Markets for the nominal bonds of different countries which are open to the residents of all countries accomplish an international allocation of the purchasing power risk of any local money supply. For example, suppose equilibrium in our open capital markets conforms to an international version of the Sharpe-Lintner model. Optimal portfolios for investors of all countries involve combining borrowing or lending at the risk-free real rate of interest with an investment in the international market portfolio, the aggregate portfolio of the invested wealths of all countries. The values of the money supplies of different countries, sold forward in the form of nominal bonds, are part of this international market portfolio. Through his holdings of nominal bonds denominated in different monies, an investor ends up with the same fraction of the total purchasing power risk of the money supply of every country.

In open, frictionless capital markets, the risk-free real rate of interest must be the same in all countries. Nominal interest rates can differ, however, because of different purchasing power risks of the monies in which the payoffs on nominal bonds are stated. We shall see that these differential purchasing power risks also work their way into the pricing of forward foreign exchange contracts, but first we must describe the market for foreign exchange.

[4]As for any other asset, the market risk of the uncertain purchasing power of the money payoff on a nominal bond can be negative. The sign and magnitude of $cov(\tilde{\pi}_{t+1}, \tilde{P}_{A,t+1})$ are, of course, an empirical matter.

A. *Purchasing Power Parity and Exchange Risk*

Let S_t^{ij} be the spot exchange rate between the monies of countries i and j at time t; it is the number of units of money i that must be given up at time t to obtain one unit of money j. Since markets for the commodity are, like capital markets, internationally open, we assume that spot exchange rates obey a strict form of purchasing power parity. The exchange rate between two monies is set so that the implied exchange rate in terms of the good is one for one. Formally,

$$S_t^{ij} = \frac{\pi_{jt}}{\pi_{it}} = \frac{l_{it}}{l_{jt}} \tag{17}$$

where π_{jt} and π_{it} are the purchasing powers of the two monies (units of the commodity that can be bought with a unit of money), while l_{it} and l_{jt} are the price levels in the two countries (units of money required to purchase one unit of the commodity). The ratio of price levels interpretation of purchasing power parity given in (17) is most common, but the ratio of purchasing powers interpretation turns out to provide a more convenient analogy to the expression developed below for forward exchange rates.

Although purchasing power parity is primarily a condition on equilibrium in open markets for the commodity, it has an important implication for portfolio opportunities. Purchasing power parity and open frictionless international capital markets are sufficient conditions for the real or commodity returns on a given security to be the same for the residents of all countries. When transformed into the monies of different countries, the nominal return on a given security can vary across countries because of changes in exchange rates. With purchasing power parity, however, the changes in exchange rates are precisely those needed to offset differential rates of change in the purchasing powers of different monies. Thus, in real terms, units of the commodity, the return on a given security is the same in all countries.

This conclusion does not depend on the assumption that there is only one consumption good. Suppose there are many consumption goods and relative prices of goods can change stochastically through time. With complete purchasing power parity, however, the relative prices of goods are the same in all countries, and the exchange rate between the monies of two countries is just the ratio of the money prices of any good in the two countries. In this situation, the return on a security can be different when transformed into the monies of different countries, but when expressed in units of any given commodity or any given linear combination of commodities (a consumption bundle), the return on a security is the same for the residents of all countries. In such a world, uncertainty about future exchange rates has no relevance for any investor's decision on how his portfolio resources should be allocated to national and international investments.

The implications of the presence or absence of purchasing power parity for the presence or absence of exchange risks from international investment are straightforward, but they are not well appreciated in the literature. For example, in analyzing the total risks from foreign portfolio investments, Michael Porter adds uncertainty about the exchange rate to the uncertainty about a foreign security's nominal return. The fact that even a little bit of purchasing power parity renders this approach invalid is never mentioned. Bruno Solnik (1976, 1977) seems somewhat aware of the issues raised by purchasing power parity, but he erroneously concludes that even with complete purchasing power parity, uncertainty about the relative prices of different consumption goods can make exchange rate uncertainty a relevant source of risk in international portfolio investments.

B. *Forward Foreign Exchange Rates*

With purchasing power parity, uncertainty about future exchange rates is irrelevant for portfolio decisions, but forward exchange rates can still contain premiums or discounts vis-à-vis predictions of future spot exchange rates. We now show that there can be a premium or discount in a forward exchange rate simply because of different purchasing power risks of the monies of the two countries.

In a forward foreign exchange contract, an

exchange rate between the monies of two countries is determined at time t, but delivery takes place at t + 1. Thus, the market must determine a forward exchange rate at time t that appropriately takes into account the risks of the future purchasing powers of the two monies. The solution to the problem can be found in the prices of the nominal bonds of the two countries, which likewise involve determining certain market values for future money payoffs with uncertain purchasing power.

Adding a subscript i to P_t in (14), in a Sharpe-Lintner international capital market, the real price at time t of a nominal bond that will pay one unit of money i at t + 1 is

$$(18)\quad P_{it} = \frac{E(\tilde{\pi}_{i,t+1}) - \phi cov(\tilde{\pi}_{i,t+1}, \tilde{P}_{A,t+1})}{1 + r_{F,t+1}}$$

This is the number of units of the commodity that must be given up at t for the right to a unit of money i to be delivered for certain at t + 1. Suppose, however, that the real price agreed at t is likewise to be delivered at t + 1. If arbitrage opportunities between the real and nominal bond markets are to be ruled out, this real forward price for a unit of money i must be

$$(19)\quad P_{it}(1 + r_{F,t+1}) = E(\tilde{\pi}_{i,t+1}) - \phi cov(\tilde{\pi}_{i,t+1}, \tilde{P}_{A,t+1})$$

From this expression for the real forward price of a unit of money, it follows that the forward exchange rate, the ratio agreed at t for exchanges of two different monies i and j at t + 1, must be

$$(20)\quad F_t^{ij} = \frac{P_{jt}(1 + r_{F,t+1})}{P_{it}(1 + r_{F,t+1})} = \frac{E(\tilde{\pi}_{j,t+1}) - \phi cov(\tilde{\pi}_{j,t+1}, \tilde{P}_{A,t+1})}{E(\tilde{\pi}_{i,t+1}) - \phi cov(\tilde{\pi}_{i,t+1}, \tilde{P}_{A,t+1})}$$

where F_t^{ij} is units of money i to be given up at t + 1 per unit of money j received. For example, suppose the real forward value of a unit of money j is twice that of a unit of money i; that is, the prices of nominal bonds denominated in the two monies tell us that the right to a unit of money j at t + 1 is worth twice as many units of the commodity as the right to a unit of money i at t + 1. Then the forward exchange rate between the two monies must be two units of money i per unit of money j.

Expression (20) for the forward exchange rate has a convenient interpretation. Expression (17) says that the spot exchange rate between two monies at time t is the ratio of their purchasing powers at t; equation (20) says that the forward exchange rate at t is the ratio of the market determined real certainty equivalents of the uncertain future purchasing powers of the two monies. The certainty equivalent of the uncertain future purchasing power of a unit of each money is just the expected value of its future purchasing power less an adjustment for its market risk, that is, the contribution of the uncertain purchasing power to the risk of the real value of aggregate invested wealth at t + 1.

To better display how any premium or discount in the forward exchange rate as a predictor of the future spot rate is linked to the risk-return characteristics of the nominal bonds of the two countries, we rearrange (20) as

$$(21)\quad F_t^{ij} = \frac{1 - \phi cov(\tilde{\pi}_{j,t+1}, \tilde{P}_{A,t+1})/E(\tilde{\pi}_{j,t+1})}{1 - \phi cov(\tilde{\pi}_{i,t+1}, \tilde{P}_{A,t+1})/E(\tilde{\pi}_{i,t+1})} \cdot \frac{E(\tilde{\pi}_{j,t+1})}{E(\tilde{\pi}_{i,t+1})}$$

In words, the premium or discount in the forward exchange rate, vis-à-vis the prediction of the future spot exchange rate given by the ratio of the expected future purchasing powers of the two monies, depends on the market risks of the two monies measured relative to their expected future purchasing powers. If the ratio of covariance to expected value is higher for money j than for money i, then per unit of expected future purchasing power, money j is relatively more risky than money i, and F_t^{ij} is less than $E(\tilde{\pi}_{j,t+1})/E(\tilde{\pi}_{i,t+1})$. Equivalently, it follows from equation (16) that the ratios of covariance to expected value in (21) determine the risk adjustments in the expected real returns on nominal bonds denominated in the two monies. The forward exchange rate F_t^{ij} is less than $E(\tilde{\pi}_{j,t+1})/E(\tilde{\pi}_{i,t+1})$ when the expected real return on the nominal bonds of country j

is higher than the expected real return on the nominal bonds of country *i*, because the nominal bonds of country *j* are riskier than those of country *i*.

Put slightly differently, a forward foreign exchange contract is an agreement to exchange the future payoffs on nominal bonds denominated in the monies of two countries. Because there can be differences in the purchasing power risks of the two monies, the real payoffs on the bonds can have different degrees of risk, in which case the bonds are priced to have different expected real returns, and the forward exchange rate is not just the ratio of the expected values of the real payoffs on the bonds. However, in our world of complete purchasing power parity a discount or premium in the forward exchange rate is in no way a consequence of an exchange risk in the system which is additional to the risks inherent in the real payoffs on the bonds. It is just another way to say that, per unit of expected real payoff, the nominal bonds of the two countries have different degrees of market risk. An analogous discount or premium would appear in the forward exchange rate for the future payoffs on two different securities, say common stocks, from the *same* national capital market.

Two papers, one by Pentti Kouri and the other by Frederick Grauer, Robert Litzenberger, and Richard Stehle, analyze models of international capital markets that are close in spirit to ours. Both papers recognize that expected returns on nominal bonds can include risk premiums because the real returns on such bonds are uncertain. And Kouri notes that the risk of the real return on a nominal bond is just the risk inherent in the underlying money supply. Both papers also recognize that there can be discounts or premiums in forward exchange rates which do not imply the existence of exchange risks that make the real returns on a given security different for the residents of different countries. And Kouri notes that the premium or discount in the forward rate is traceable to the differential risks of the money supplies of the two countries.

However, although its function is implicit in their analyses, Kouri and Grauer, Litzenberger, and Stehle do not bring out the role of purchasing power parity in rendering exchange rate uncertainty irrelevant for portfolio decisions. Moreover, neither paper gives any attention to the economics of money or even how money gets into the system, so one does not get a clear picture of how the purchasing power risk of the money supply works its way into the pricing of nominal bonds and into the setting of forward foreign exchange rates. Our analysis of the roles of money and nominal bonds is meant to make such insights more accessible.

VI. Summary

This paper studies the role of money as a durable productive asset and the role of nominal bonds in transforming money into a portfolio asset. Our major point is that nominal bonds allow current users of money to sell its future services forward. In this way nominal bonds allow an allocation of the current services of the money supply according to its current productivity and without regard for its future risk and an allocation of the future purchasing power risk of the money supply according to the willingness of investors to bear this risk in exchange for return and without regard for investor needs for the future transactions services of money. In international capital markets, nominal bonds allow the purchasing power risk of a country's money supply to be borne by the residents of other countries even though the money might provide transactions services only within the borders of the country issuing it.

REFERENCES

E. F. **Fama**, "Multiperiod Consumption-Investment Decisions," *Amer. Econ. Rev.*, Mar. 1970, *60*, 163–74.

S. **Fischer**, "The Demand for Index Bonds," *J. Polit. Econ.*, June 1975, *80*, 509–34.

———, "Money and the Production Function," *Econ. Inquiry*, Dec., 1974, *12*, 517–33.

M. **Friedman**, "The Quantity Theory of Money: A Restatement," in his *Studies in the Quantity Theory of Money*, Chicago 1956.

F. L. A. Grauer, R. H. Litzenberger, and R. E.

Stehle, "Sharing Rules and Equilibrium in an International Capital Market under Uncertainty," *J. Finan. Econ.*, June 1976, *3*, 233–56.

P. J. K. Kouri, "International Investment and Interest Rate Linkages under Flexible Exchange Rates," in Robert Z. Aliber, ed., *The Political Economy of Monetary Reform*, New York 1977.

J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *45*, 13–37.

N. Liviatan and D. Levhari, "Risk and the Theory of Indexed Bonds," *Amer. Econ. Rev.*, June 1977, *67*, 366–75.

M. Porter, "A Theoretical and Empirical Framework for Analyzing the Term Structure of Exchange Rate Expectations," *Int. Monet. Fund Staff Papers*, Nov. 1971, *18*, 613–42.

W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

B. H. Solnik, "Sharing Rules and Equilibrium in an International Capital Market: A Comment," Centre D'Enseignement Supérieur des Affaires, work. paper, Jouy-en-Josas, France, 1976.

———, "Testing International Asset Pricing: Some Pessimistic Views," *J. Finance*, May 1977, *32*, 503–12.

J. Tobin, "Liquidity Preference as Behavior Towards Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–85.

Testing the Demand Side of Comparative Advantage Models

By YUTAKA HORIBA*

Both the Ricardian theory of comparative advantage and the more modern version often associated with the names of Eli Heckscher and Bertil Ohlin (hereafter H-O) have been subjected to many empirical tests. But the results of the tests have been mixed, and on the whole, inconclusive.[1] While the Ricardian theory attributed the source of pretrade commodity price differences among nations to broadly specified production conditions as represented by factor- (labor) productivity ratios, the H-O theory in its simpler version attributed the same to relative differences in factor endowments prevailing among different nations. The two theories are broadly similar, however, in the sense that they both identify the ultimate source of comparative advantage on the supply side. No crucial role is assigned by either theory to the demand side in the determination of comparative advantage.[2] Indeed, the demand side is effectively neutralized, at least in the standard H-O model, by the strategic and commonly made assumption that consumer preferences are internationally identical and homothetic.[3] But how crucial is such an assumption, particularly when there are many commodities with which almost all empirical tests of comparative advantage are concerned? Can the "robustness" of this assumption be established, and hence its empirical relevance, short of directly testing the assumption itself? The existing literature has not dealt explicitly with these questions in the context of the comparative advantage issue.

There have also been other difficulties associated with testing the respective theories

*Associate professor of economics, Tulane University. I wish to acknowledge useful comments and suggestions I received from Martin Bronfenbrenner, Robert Stern, and the managing editor of this *Review* on earlier versions of the paper. I also gained from the comments of seminar participants at Tohoku University, the Economics Institute, and the Stockholm School of Economics. I alone remain responsible for any errors.

[1]As is now widely known, the pioneering work of testing the Ricardian theory of trade was undertaken by G.D.A. MacDougall (1951). This was followed, with substantively the same methodology, by Stern (1962), Bela Balassa, MacDougall et al., Mordechai Kreinin, and more recently with a somewhat different methodology, by Herbert Glejser. Kreinin's study is particularly suggestive, however, in that the remarkably stable empirical relationships which were found in the earlier studies with American and British data do not extend to several other countries which he considered. The first systematic attempt to test the H-O theory was due to Wassily Leontief, whose well-known finding on the *U.S.* capital position caused a major controversy. (See Robert Baldwin, for example, for a perspective review of the Leontief Paradox.) Leontief's work was followed, with widely mixed results, by Masahiro Tatemoto and Shinichi Ichimura, Donald Wahl, Ranganath Bharadwaj, Karl Roskamp, Michael Hodd, and Roskamp and Gordon McMeekin, among others, all of whom applied the input-output technique for observations from additional countries. More recently, however, a multivariate approach to testing the H-O theory, with diverse empirical specifications, has been undertaken. See, for example, Gary C. Hufbauer, William Branson and Helen Junz, Baldwin, Edward Leamer, Jon Harkness and John Kyle, Stern (1976), Branson and Nikolaos Monoyios, and Harkness.

[2]Ohlin (especially ch. 1) emphasized the interplay of demand and supply conditions in the determination of commodity and factor prices. The demand side appeared in many of Ohlin's passages, but it was never given as critical an importance as the supply side in his analysis of the cause of trade.

[3]Consumer preferences are said to be homothetic if indifference hypersurfaces have the same gradient for any consumption vector unique up to a scalar. The direct implication of homothetic preferences is that income elasticities of demand are identically unitary for all goods. It is of some interest to note, however, that John Stuart Mill also made a similar assumption in demonstrating his law of international values in a two-commodity context (ch. 18). As pointed out by John Chipman (pp. 484–85), Mill's assumptions (p. 155) were tantamount to positing a utility function (internationally identical) of the form, $log\ U = \delta_1\ log\ C_1 + \delta_2\ log\ c_2$, where the c_i's denote the units of consumption, and the α_i's are a set of parameters denoting consumption expenditure shares. The assumption of internationally identical constant consumption expenditure shares, often extended over many goods, was also employed in modern works on the classical theory of trade. See, for example, Thomas Whitin and Lionel McKenzie.

of comparative advantage. The earlier method of testing the Ricardian theory, for instance, employed regression analysis which specified for the dependent variable the relative share of exports of a pair of countries in third markets of the world. In this context the independent variable was often given by either intercountry labor-productivity ratios, or the corresponding unit labor cost ratios. Jagdish Bhagwati (1964) argued, however, that there is no theoretical reason why there should exist any simple and monotonic transformation of the labor-productivity ratio into the export share across industries.[4] If this criticism is valid, then it appears that a similar argument may also apply against a number of different regression specifications which have been attempted for testing the H-O theory.

If we now return to either theory of comparative advantage, we note that there is a presumption that comparative advantage stemming from the supply side is revealed by the observable *post*trade pattern of production under conditions of free trade. In the two-country, two-commodity Ricardian model, for instance, with fixed input usage (homogeneous labor) per unit production, the resource allocation after trade takes the extreme form of complete specialization, at least for the smaller country. Likewise, in the standard H-O model with two goods, the home country will produce after an opening of trade proportionately more of the good in which it has comparative advantage than what its trading partner will produce. Such a pattern of resource concentration after trade is associated with a bias in the relative shapes of the respective country's production possibility curves. In particular, the "supply bias" may be defined such that for any common price vector the proportion of goods produced at equilibrium is always higher in one country. In the two-commodity Ricardian model with complete specialization, the proportion of goods produced tends to be either zero or infinity, but the above characterization of supply bias still applies for any common price ratio bounded by the respective country's autarky price ratios.

This paper is motivated to demonstrate: 1) there are circumstances in which the hierarchy of comparative advantage defined in terms of pretrade interregional (or intercountry) price differentials may be captured in the output space as posttrade output proportions produced in the multicommodity context; and 2) there are some explicit and testable implications of demand, and in particular, the assumption of identical and homothetic preferences, on the transformation of output proportions into the corresponding pattern of trade. For the purpose of elucidating the latter linkage, I construct a broadly specified structural model with many goods. I then test the model, using the *U.S.* interregional manufacturing trade and production data as the basis of my empirical observations. It is apparent that policy distortions are largely absent in the course of interregional trade, and we are able to draw sharper conclusions concerning the role of demand in the determination of the pattern of trade. The choice of the *U.S.* regions for this purpose has been facilitated by the recent availability of a consistent set of data on commodity trade flows among the *U.S.* regions. Therefore, the term "region" is used in much of what follows, but it can be replaced without much difficulty by "country" as in the usual context.[5]

I. The Model

Following the traditional comparative advantage models, assume that competitive equilibrium prevails everywhere with free trade, and that no transportation cost is incurred in the movement of goods. We consider, for simplicity, an open economy at

[4]There is, however, some controversy over Bhagwati's objection. See the exchange of views among Donald Daly, Balassa, and Bhagwati, in Daly's book.

[5]The distinction between regions and nations is mainly superficial from the standpoint of the pure theory of comparative advantage, with obvious qualifications concerning the differing extent of factor mobility, transport costs, policy impact, etc. It may even be said that the H-O theory proper, in its pristine form, is more readily applicable to interregional trade, as Ohlin himself began with an analysis of interregional trade, and only then extended the analysis with a series of modifications to international trade.

posttrade equilibrium, where all external as well as internal equilibrating processes have already taken place. For what follows, however, we depart from the spirit of the Ricardian theory somewhat (but not necessarily from that of the H-O theory) in assuming that each trading region continues to produce after trade some positive quantity of each traded good, barring complete specialization.[6]

Let q_i and Q_i $(i = 1, \ldots, n)$ denote the units of the ith output produced in a given region and in its trading region, respectively. For convenience, the lowercase letters refer to the given (home) region, and the uppercase letters refer to its trading partner. The partner region in the multiregional empirical context to follow is evaluated by the combination of all outside regions, and the two-region formulation is retained throughout. By our assumption of common production, we have

$$(1) \qquad q_i > 0 \qquad i = 1, \ldots, n$$

Supply bias in the higher dimension can now be defined as follows. We say that supply bias occurs if for any common price vector it is possible to number commodities such that

$$\frac{q_i}{q_j} > \frac{Q_i}{Q_j} \qquad \text{for all } i \text{ and } j \quad (i, j = 1, \ldots, n; i < j)$$

where each quantity denotes the equilibrium output associated with the given price vector. In particular, this implies the following chain of inequalities at the posttrade equilibrium situation (designated by an asterisk) with interregionally equalized commodity prices:

$$(2) \qquad \frac{Q_1^*}{q_1^*} < \frac{Q_2^*}{q_2^*} < \ldots < \frac{Q_n^*}{q_n^*}$$

It is possible to relate that above chain of output ratios to the pretrade commodity price ratios in the absence of demand bias in the following sense. Let p_i' and P_i' denote the pretrade equilibrium prices of the ith output in the respective regions. We say that demand bias occurs if for some pair of goods, say i and j, an inverse relationship between their pretrade equilibrium commodity price ratio and the corresponding output ratio is broken such that $p_i'/p_j' > P_i'/P_j'$ for some i and j $(i < j)$, when $q_i/q_j > Q_i/Q_j$ for any common price vector. Hence, in the absence of such demand bias, the chain of inequalities in (2) necessarily corresponds to the hierarchy of comparative advantage expressed in terms of the pretrade relative commodity prices (i.e., unit costs of production under perfect competition):[7]

$$(3) \qquad \frac{P_1'}{p_1'} > \frac{P_2'}{p_2'} > \ldots > \frac{P_n'}{p_n'}$$

Let e_i denote the regional trade flow of the ith output from the home region. A positive value of e_i indicates an outflow from the home region, and the negative value indicates an inflow into the same region. Then, at posttrade equilibrium, we must have

$$(4) \qquad e_i^* = q_i^* - c_i^*$$

where c_i^* denotes the units of the given good consumed in the home region. Since any outflow from the home region is necessarily an inflow into the partner region in the two-region framework, we also have $e_i^* = -E_i^* = -Q_i^* + C_i^*$.

On the consumption side, we have generally,

$$(5) \qquad c_i^* = k_i^* \, (q_i^* + Q_i^*)$$

where k_i^* denotes the equilibrium proportion

[6]In reality, this assumption is upheld in the majority of manufacturing trade among the *U.S.* regions, at least at the level of industrial classification used in this paper. One possible theoretical justification of this assumption (as opposed to a justification based on casual empiricism) is an explicit recognition that employed labor typically embodies an extensive job-specific training, which may not be readily transferable among different industries. We may then assume, as the simplest case, that instead of only one type of labor as posited by the Ricardian theory, there are n different types of labor (or any other factor for that matter), each specific to the given industry at least in the short run. Hence, if we assume fixed production coefficients, the efficient production set is characterized by an n-dimensional vector point rather than an n-dimensional hyperplane of the strictly Ricardian variety. In such a case, though extreme, the pattern of pretrade nonspecialization will be strictly preserved after trade.

[7]The concepts of demand and supply bias are offered here in the definitional sense. Nonetheless, it is possible to find a set of theoretical (sufficient) conditions under which inequalities (2) and (3) necessarily hold. See Appendix A.

of the ith output consumed in the home region relative to the aggregate quantity produced in both regions combined. The values of k_i^* are in general determined by demand conditions; hence, equation (5) represents a broad specification of home consumption which is associated with the posttrade equilibrium prices and income.

Finally, we assume an equilibrium balance of payments:

$$(6) \qquad \left(\sum_i P_i^* e_i^*\right) + T = 0$$

where P_i^* denotes the interregionally equalized posttrade unit price of the ith output, and T denotes net capital inflow into the home region, offsetting the commodity trade imbalance (if any).[8]

II. Some Testable Implications of the Model

Assume now that aggregate demand is regionally identical with homothetic preferences. It then follows that the aggregate equilibrium consumption vectors are strictly proportional between the regions, given the consumers in the two regions face an identical set of equilibrium commodity prices. Therefore, we can write

$$(7) \qquad c_i^* = \mu C_i^* \qquad i = 1, \ldots, n$$

where μ denotes the fixed factor of proportionality. We thus obtain

$$(8) \quad c_i^* = \frac{\mu}{1+\mu}(c_i^* + C_i^*) = \frac{\mu}{1+\mu}(q_i^* + Q_i^*)$$

where we made use of the identity between total consumption $(c_i^* + C_i^*)$ and production $(q_i^* + Q_i^*)$ for the two regions combined. Hence, equation (5) can be rewritten as

$$(9) \qquad c_i^* = b\,(q_i^* + Q_i^*)$$

where $b = \mu/(1+\mu)$.

Replacing c_i^* in (9) by $q_i^* - e_i^*$ from equation (4), and applying equation (6), we find

$$(10) \quad b = \left(\sum P_i^* q_i^* + T\right) \Big/ \left(\sum P_i^* q_i^* + \sum P_i^* Q_i^*\right)$$

Dividing equation (9) by q_i^* (>0), we obtain

$$(11) \qquad y_i^* = (1-b) - b(Q_i^*/q_i^*)$$

where $y_i^* = e_i^*/q_i^*$, denoting the ith trade flow from the home region scaled by the level of home production of the given output.

Let the commodities be numbered such that output ratios, Q_i^*/q_i^* $(i = 1, \ldots, n)$, follow the ranking as given by (2). But equation (11) implies, in turn, that

$$(12) \qquad e_i^* \gtreqless 0 \text{ as } Q_i^*/q_i^* \lesseqgtr \pi$$

$$\text{where} \qquad \pi = \frac{\left(\sum P_i^* Q_i^*\right) - T}{\left(\sum P_i^* q_i^*\right) + T}$$

Hence, a good is exported (imported) as its output ratio (Q_i^*/q_i^*) is greater (less) than the aggregate production ratio (π) adjusted for the trade imbalance.

We thus obtain the following summary of the preceding results. If aggregate demand is identical between regions with homothetic preferences, then there exists a linear relationship between trade flow scaled by local production (y_i^*) and the corresponding output ratio (Q_i^*/q_i^*) such that

(a) The ranking of the scaled trade flow inversely corresponds to the ranking of the output ratios.

(b) The direction of trade is determined by the relative magnitude of output ratios in relation to the adjusted aggregate production ratio as given by (12).

(c) The slope of the linear relationship between the scaled trade flow and the output ratios is the aggregate production proportion (corrected for trade imbalance) as given by (10).

[8]See Harry Richardson (pp. 262–70), for example, for further reference and discussion concerning the nature of regional balance of payments and its adjustment mechanism. We assume that an overall equilibrium in the balance of payments prevails in the regional context. As we shall see in Section III, however, this assumption is not crucial to the major conclusions obtained in this paper.

III. Statistical Results

The three main implications of the model obtained in the preceding section are testable, and we now turn our attention in this direction. The basic data used for this purpose are the manufacturing commodity trade flows compiled for the multiregional input-output model (*MRIO*) of the United States. The data have been obtained from computer tape distributed by the National Technical Information Service, U.S. Department of Commerce. In particular, the data used in this study are for the fifty-one *MRIO* manufacturing industries (*MRIO* industry No. 14–64) for 1963 among the nine *U.S.* census regions.[9]

Consistent with the specification of the model, each census region is paired with the composite of all other regions. Following equation (11), the dependent variable is measured by the commodity flow from one region to all other regions, scaled by the level of output produced in the given region. Since there are simultaneous regional trade inflows and outflows for the majority of goods, the scaled trade flow is measured in net terms as outflow minus inflow.[10] The interregional commodity trade flow is evaluated in gross value of shipments. Detailed data on the breakdown of these shipments according to final demand are not available, and only the gross value of each output consumed in the destination region can be computed from the available data, as the difference between the gross value of production and the gross value of shipments. We henceforth assume that some constant fractions of interregional commodity shipments (as well as gross production) are for final consumption by households. These fractions clearly differ in magnitude among different goods, depending upon the nature of each good and its classification. However, the major statistical results to follow would not be affected by the net valuation basis (had it been available), as long as the assumed factors of proportionality are identical among regions for comparable goods.[11]

The statistical test results are presented in summary forms in Table 1. To remove the potential influence of extreme values of the scaled net trade ratio on the statistical tests performed, I applied different deletion criteria on the data. The criteria pertain to the extent of self-sufficiency of each region in meeting its gross consumption needs. Clearly, if the scaled net trade flow assumes a large negative value (say, -9), the corresponding self-sufficiency ratio of consumption to intraregional production is necessarily small (.10 in the example). It turned out, however, that the results, obtained under different deletion criteria ranging from 10 to 50 percent self-sufficiency, are in all essential aspects remarkably similar. In Table 1, only the results of the more stringent 40 percent self-sufficiency criterion are reported.

Predictive statements (a) and (b) drawn in the previous section are nonparametric in nature, and we begin by testing these implications. To test the rank-ordering association between the scaled trade flow and the output ratios, I computed the Spearman rank-correlation coefficient for each census region. As reported in the table (col. (6)), the rank correlation coefficients are negative as postulated, and they are highly significant in all nine cases.[12] Hence, this finding constitutes a strong empirical support for the first implication of homotheticity on the pattern of trade. Although there is considerable variation in the actual (gross) consumption proportions

[9]See John Rodgers and Karen Polenske for detailed descriptions of the *MRIO* model and the data. I am indebted to Michael Ray for programming assistance in the retrieval and tabulation of data from the tape.

[10]The occurrence of simultaneous inflow and outflow may be attributed to the level of aggregation used in industrial classification. See Herbert Grubel and Peter Lloyd (especially chs. 5–7), however, for an attempt to explain this phenomenon on other (economic) grounds.

[11]Let $e_i^* = \alpha_i e_i^f$, $q_i^* = \alpha_i q_i^f$, and $Q_i^* = \alpha_i Q_i^f$, where e_i^f, q_i^f, and Q_i^f are in gross values, and e_i^*, q_i^*, and Q_i^* are in the corresponding values destined for final consumption with the factor of proportionality given by α_i $(i = 1, \ldots, n)$ Substituting these into equation (11), we find that the α_i's cancel for each output on both sides of equation (11).

[12]The null hypothesis of no correlation is rejected in each case at the 1 percent level or higher. It should be noted that significance tests of rank correlation are possible only with respect to the null hypothesis. In particular, it is not possible to test the hypothesis of perfect rank correlation.

TABLE 1—STATISTICAL SUMMARY

U.S. Census Region[c]	Proportion of Adjusted Aggregate Production[d] (1)	Unconstrained Regression[e] Constant (2)	Slope (3)	R^2 (4)	Constrained Regression Slope[f] (5)	Spearman Rank Correlation Coefficient (6)	*Chi*-Square Statistic[g] (7)	Mean Absolute Percent Deviation[h] (8)	Number of Observations[i] (9)
South Atlantic	.127	.490[b]	−.062[b]	.723	−.093[b]	−.841[b]	24.4[b]	30.1	41
		(.087)	(.006)		(.004)	(.087)			
East South Central	.049	.571[b]	−.026[b]	.748	−.038[b]	−.891[b]	5.6[a]	22.1	44
		(.070)	(.002)		(.002)	(.070)			
West South Central	.082	239[b]	− 027[b]	686	−.049[b]	−.867[b]	7.7[b]	32.1	36
		(.087)	(.003)		(.003)	(.085)			
New England	.067	.504[b]	−.030[b]	.808	−.049[b]	−.898[b]	18.6[b]	22.9	46
		(045)	(.002)		(.002)	(.066)			
Middle Atlantic	.214	.500[b]	− 101[b]	.763	− 179[b]	−.835[b]	12.6[b]	15.9	48
		(.047)	(.008)		(007)	(.081)			
East North Central	.253	.673[b]	−.146[b]	.811	−.210[b]	−.801[b]	5.5[a]	22.4	48
		(047)	(.010)		(007)	(.088)			
West North Central	.076	.512[b]	−.042[b]	.696	− 062[b]	− 861[b]	22.0[b]	19.6	45
		(.095)	(.004)		(.002)	(.077)			
Mountain	.033	426[a]	−.017[b]	656	−.026[b]	− 828[b]	4.3[a]	29.0	19
		(164)	(003)		(002)	(.136)			
Pacific	.134	.099	−.044[b]	422	− 103[b]	− 823[b]	4.3[a]	22.5	46
		(107)	(008)		(006)	(086)			

Note Numbers in parentheses are standard errors

[a]Significant at the 5 percent level

[b]Significant at the 1 percent level

[c]South Atlantic includes Delaware, Maryland, District of Columbia, Virginia, West Virginia, North Carolina, South Carolina, Georgia, and Florida. East South Central includes Kentucky, Tennessee, Alabama, and Mississippi. West South Central includes Arkansas, Louisiana, Oklahoma, and Texas. West North Central includes Minnesota, Iowa, Missouri, North Dakota, Nebraska, Kansas, and South Dakota. Middle Atlantic includes New York, New Jersey, and Pennsylvania. East North Central includes Ohio, Indiana, Illinois, Michigan, and Wisconsin. New England includes Maine, New Hampshire, Vermont, Massachusetts, Rhode Island, and Connecticut. Mountain includes Montana, Idaho, Wyoming, Colorado, New Mexico, Arizona, Utah, and Nevada. Pacific includes Washington, Oregon, California, Alaska, and Hawaii.

[d]Evaluated as $(\Sigma P_i^* q_i^* - \Sigma P_i^* e_i^*)/(\Sigma P_i^* q_i^* + \Sigma P_i^* Q_i^*)$. (See the text.) This computation applies to the included industries only; hence, the sum of this column is not precisely unity.

[e]See equation (13) for regression specification.

[f]The regression slope is estimated from the constrained minimization of $\Sigma(Y_i - 1 - \hat{b}X_i)^2$, where Y_i and X_i denote the dependent and independent variable, respectively, as given by (13). From the first-order condition we obtain $\hat{b} = (\Sigma Y_i X_i - \Sigma X_i)/\Sigma X_i^2$ Considering the variance of $\hat{b}$, we obtain $var(\hat{b}) = \sigma^2/\Sigma X_i^2$, where σ^2 denotes the variance of the random disturbance term. Hence, applying $\hat{\sigma}^2 = \Sigma(Y_i - 1 - \hat{b}X_i)^2 / (n - 1)$ in this case, we find the estimate of $var(\hat{b})$ as $\hat{\sigma}^2/\Sigma X_i^2$, from which the standard error is obtained as its square root.

[g]See the text for the formula used.

[h]Evaluated as $100\ \Sigma_i 2[|c_i - c_i^*|/(c_i + c_i^*)]/n$, where c_i denotes the actual consumption, and c_i^* denotes the consumption level computed from (9) with the b-value as given in col. (1).

[i]The deletion criterion applied to the data is that the given region be at least 40 percent self-sufficient in meeting its consumption needs, i.e., $y_i^* \geq -1.5$.

across commodities, evidently it is dominated by a variation in output proportions on the production side so as not to upset the rank-ordering association between trade and production.[13]

[13]It is possible to relax the assumption of strict homotheticity somewhat and obtain a more general condition under which this type of rank-ordering association holds. See Appendix B.

Let us now consider the statement (b) entailing the direction of trade. Following the relationship (12), a two-by-two contingency table can be tabulated, bearing the sign of net trade on the one hand, and the sign of output ratio net of the adjusted aggregate production ratio (π) on the other. It is then possible to apply the well-known *chi*-square test of independence to the actual sign frequency distribution. In particular, the test statistic is given

by

$$\frac{n\left(|a_{11}a_{22} - a_{12}a_{21}| - \frac{n}{2}\right)^2}{(a_{11} + a_{12})(a_{21} + a_{22})(a_{11} + a_{21})(a_{12} + a_{22})}$$

where a_{ij} is the observed frequency in the ith row (denoting the sign of net trade) and the jth column (denoting the sign of output ratio less π), and n stands for the number of observations. The degree of freedom associated with this statistic is one. As reported in the table (col. (7)), the computed *chi*-square values fall in the 1 percent critical region in five cases, and in the 5 percent critical region in the remaining four cases.[14] Consequently, the assumed independence of the direction of trade can be rejected, lending further support to the nonparametric implications of the model.

Consider now a direct test of the linear relationship as obtained in equation (11). Since the parameter b enters into both the intercept and the slope term, we have an apparent identification problem for the purpose of estimating b from regression analysis. This, however, can be circumvented by a slight rearrangement of equation (11) as follows:

$$(13) \quad y_i^* = 1 - bx_i^* + u_i \qquad i = 1, \ldots, n$$

where $x_i^* \equiv (Q_i^* + q_i^*)/q_i^*$, and u_i denotes the random disturbance term, assumed to be normally distributed with zero mean, constant variance, and zero covariance everywhere. The unconstrained ordinary least squares estimation result reveals a high degree of clustering of the data points in each case as evidenced by the value of R^2. As can be verified in the table, however, the separately computed aggregate production proportion corrected for trade imbalance (col. (1)) falls below any reasonable (say, 99 percent) confidence interval constructed on the basis of the regression slope and its standard error. Likewise, the theoretical unit value falls outside any reasonable confidence interval constructed around each regression intercept, exceeding the respective upper bounds by a substantial margin relative to the size of standard error. Hence, the more stringent parametric implications of homotheticity appear to be contradicted by these observations.

An alternative regression estimation, which is based on constraining the fitted intercept to be unity, does not reverse the above finding. As can be seen in the table (col. (5)), the estimated slope under this specification increased in absolute value in each case towards the respective adjusted aggregate production proportion, However, the latter continues to fall short of any acceptable confidence interval constructed on the basis of the former.

It may be argued that the above discordance with the statement (c) is a possible outcome of disequilibrium in the regional balance of payments. But this will not affect our conclusion in the present case, inasmuch as the adjustment factor (T), typically small in relation to the gross value of production, has been measured by the commodity trade imbalance in the first place. If there is disequilibrium in the balance of payments, equation (6) is replaced by

$$(\sum P_i^* e_i^*) + T = \sigma$$

where σ measures the extent of imbalance. Consequently, equation (10) becomes

$$b = (\sum P_i^* q_i^* - \sum P_i^* e_i^*) / (\sum P_i^* q_i^* + \sum P_i^* Q_i^*)$$

But this is precisely how the b-value is separately computed in the present case. In addition, inclusion of the excluded industries in the separate computation of the b-value did not affect the reported result to any significant degree.

The manner in which the regression line deviates in each case from the hypothetical line may suggest that there are some unaccounted for factors which serve to depress the actual trade flow below the predicted level. Suppose, for example, that transport costs "sacrifice" trade in proportion to its volume. Thus, we may have a simple auxiliary relationship,

[14] In the case of the Mountain region for which the number of observations is less than twenty, statistical significance was established with use of the Fisher test. See Sidney Siegel, pp. 96–111, for this method.

$$e_i^t = \beta e_i^* \qquad 0 < \beta < 1 \tag{14}$$

where e_i^t denotes the trade flow with transport costs, and e_i^* denotes as before the trade level associated with homothetic preferences without transport cost, whence the factor of proportionality of the sacrificed trade is given by $1 - \beta$. Substituting equation (14) into equation (11), we find that β serves to alter both the intercept and the slope term in the downward direction as may be desired.[15]

Clearly, other conjectures appear to be equally plausible, such as imperfect markets and the role of uncertainty in reducing the level of trade from what would have prevailed otherwise. However, it is difficult to resuscitate the strict parametric implications of homotheticity on these grounds. Indeed, we find that the actual (gross) consumption points deviate rather substantially for some sectors from the implied homothetic level. For the purpose of this comparison, the consumption level implied by homothetic preferences can be evaluated directly from equation (9) for each output. A simple average of the absolute percentage deviations observed between the actual consumption and the corresponding implied consumption level is tabulated for each region in the table (col. (8)). These numbers are presented as a summary indication of the empirical magnitudes of the percentage deviations involved.[16] It seems safe to conclude, therefore, that the empirical thrust of the traditional assumption of homotheticity on the demand side is restricted mainly to its nonparametric implications concerning the direction of trade and its rank-ordering association with the pattern of production in the multicommodity context.

IV. Conclusions

It is well known that if Ricardo's one-factor (labor), two-commodity formulation of comparative advantage (ch. 7) is generalized to include many commodities, there will be a chain of goods ranked according to intercountry labor-productivity ratios such that any commodity exported by a given country has always a higher relative labor productivity than any imported commodity.[17] Similarly, in the H-O trade model with two factors (capital and labor) and many commodities, it is possible to arrange goods according to capital-labor ratios employed in production, and the chain of goods thus arranged reflects the hierarchy of comparative cost advantage.[18] The role of demand is decidedly in the background in both theories of comparative advantage, as demand conditions are invoked primarily to indicate where the chain of goods is broken into exportable goods on one hand and importable goods on the other.

I have argued in this paper that within a framework of broadly specified structural relations for an open economy, characterized by general equilibrium and diversification of productive activities, there are explicit and testable implications of the homotheticity of identical consumer preferences as they pertain to the pattern of trade in a multicommodity situation. In particular, homotheticity implies that the trade flow scaled by production is a monotonic transformation of output proportions, and that the latter in turn determine the direction of trade in relation to the relative size of aggregate production. In this context, it is possible to relate the observed output production ratios to the structure of pretrade price differentials under the generalized notion of supply bias and in the absence of demand bias in the higher dimension. This is in keeping with the general presumption that comparative advantage, if unfettered by policy and other barriers to trade, should manifest itself in the extent of posttrade resource concentration.

Observations from the *U.S.* interregional

[15]Note that the particular manner in which transport costs are introduced leaves both the rank correlation and the direction of trade unaffected. The predicted direction of trade is the same as before inasmuch as the point of intersection of the regression line with the horizontal axis remains the same.

[16]Consistent with the other columns of the table, the mean absolute percentage deviation is evaluated for each region from the included outputs only. Hence, the outputs included for computation are not uniform among regions.

[17]See Gottfried Haberler, pp. 136–40, for example, on this point.

[18]See Ronald Jones (pp. 5–6). As pointed out by Bhagwati (1972), however, this proposition holds only in the absence of factor-price equalization.

trade in manufacturing have produced strong evidence in support of the nonparametric implications of homotheticity, even though its parametric implications on trade are contradicted by the evidence presented. Questions concerning the proper definition of a region as well as other data limitations always remain. One assumption which was necessitated by the lack of data concerns fixed proportionality between gross values of interregional shipments and the final consumption in the region of destination of these shipments. However, the assumed identity of proportionality extends only across regions for each separate output, and the factors of proportionality are allowed to be different among different outputs. Hence, this does not appear to detract in a crucial way from the thrust of the empirical findings of Section III, though the validity of this assumption can only be tested by more detailed data on the breakdown of interregional shipments according to the components of final demand.

Attention has been focused in this paper on the more practical aspect of the generally neglected demand side in the context of the comparative advantage issue. In the process any attempt to test the exact *source* of comparative advantage (be it Ricardian or of the H-O variety) arising on the supply side has been avoided altogether. It is now apparent, however, that the latter issue should be sought as much in the locational aspect of production as in the actual pattern of trade. Indeed, it is surprising that although Ohlin, for instance, placed his major emphasis on location in the analysis of international and interregional trade (see, especially, chs. 10–13), none of the international tests of the H-O theory cited in this paper examined this aspect of comparative advantage. The inevitable implication of the largely mixed empirical results of international tests has been that comparative advantage based on either theory is not a regularly observed economic phenomenon. However, since the pattern of international trade is evidently quite sensitive to policy distortions (see William Travis, for example), the negative implication on the comparative advantage theories would indeed be more convincing (or, some seemingly "good" findings less convincing) if comparative advantage of either variety also fails to register empirically as a major supply determinant of the international pattern of resource allocation and production.

Appendix A: Demand and Supply Bias in the Higher Dimension— An Illustration

The concept of supply bias in the text, together with that of demand bias, extends the two-commodity case in an obvious and perhaps the simplest manner. Although these concepts were offered in the definitional sense, it is possible to find a set of conditions under which inequalities (2) and (3) necessarily hold. This section is intended to illustrate some of these conditions, and to demonstrate the nonvacuous nature of these concepts.

Suppose, for example, that we have industry-specific fixed factor coefficients of production as in footnote 6, with zero marginal rates of factor substitution everywhere. Then, the chain of inequalities in (2) is equivalent to

$$(S_1/s_1)(a_1/A_1) < (S_2/s_2)(a_2/A_2) < \dots < (S_n/s_n)\,(a_n/A_n)$$

where s_i and S_i denote the respective region's industry-specific factor supplies (fully employed), and a_i and A_i denote the corresponding fixed input coefficients. Hence, supply bias in our sense occurs if (a) the relative factor supplies are given by the ranking, $S_1/s_1 < S_2/s_2 < \dots < S_n/s_n$, with regionally identical coefficients of production; or (b) factor-productivity differences exist among regions and are given by the ranking, $a_1/A_1 < a_2/A_2 < \dots < a_n/A_n$, with regionally similar (i.e., identical up to a scalar) distribution of the different factor supplies. Case (a) may be considered as a special case of the H-O model; likewise, case (b) may serve as a variant of the Ricardian model, in which the given string of inequalities refers to the hierarchy of comparative advantage defined in terms of labor productivity differences.

Consider now an aggregate utility function

assumed to be identical among regions, $U = U(c_1, c_2, \ldots, c_n)$, where U is quasi concave and differentiable. We have at pretrade equilibrium $c_i' = q_i'$ in each region, and $p_i'/p_j' = MRS_{ij}$, where $MRS_{ij} = (\partial U/\partial c_i) / (\partial U/\partial c_j)$, evaluated at $c_i = c_i'$ and $c_j = c_j'$ for all i and j. Note that demand bias in the higher dimension can be ruled out on theoretical grounds if $MRS_{ij} = f_{ij}(c_i'/c_j')$ for all i and j $(i \neq j)$ in which $f'_{ij} = dMRS_{ij}/d(c_i'/c_j') = dMRS_{ij}/d(q_i'/q_j') < 0$, so that the pretrade equilibrium price ratio p_i'/p_j' is a monotonically decreasing function of q_i'/q_j' only. Consider, for example, a utility function of the Millian form given in footnote 3, extended to the n-commodity case. Here, $MRS_{ij} = (\alpha_i/\alpha_j)(c_i'/c_j')^{-1}$; hence, $dMRS_{ij}/d(q_i'/q_j') < 0$. Likewise, it can be verified that a utility function of the CES form, $U = (\Sigma\lambda_i c_i^{-\theta})^{-1/\theta}$ in which $\lambda_i > 0$ for all i, rules out the demand bias. However, not all homothetic utility functions satisfy the above condition. For example, consider $U = (W)^{1/2}$, where $W = \Sigma_i\sigma_i c_i^2 + \Sigma_i\Sigma_j\sigma_i\sigma_j c_i c_j, \sigma_i > 0$ for all $i, i \neq j$. Clearly, U is homothetic. But $MRS_{ij} = (\sigma_i c_i + \Sigma_r\sigma_i\sigma_r c_r)/(\sigma_j c_j + \Sigma_r\sigma_j\sigma_r c_r)$, and the simple monotonicity is no longer guaranteed. I emphasize, however, that the above condition of monotonicity with separability constitutes merely a sufficient condition for the nonoccurrence of demand bias in the higher dimension.

APPENDIX B: RELAXATION OF THE STRICT HOMOTHETICITY OF PREFERENCES

It is possible to relax the assumption of strict homotheticity somewhat without violating the resultant rank-ordering association between the scaled trade flow (y_i^*) and output ratios (Q_i^*/q_i^*). Let $z_i^* = Q_i^*/q_i^*$, and $x_i^* = (q_i^* + Q_i^*)/q_i^*$; hence, x_i^* denotes the posttrade inverse share of production of the ith output in the given region relative to the national production. Replace the constant consumption proportion b in equation (11) by k_i^* from equation (5), in order to allow for nonhomothetic preferences. We now obtain the following.

PROPOSITION: *Assume that income elasticities of demand are nearly identical among goods in the sense that for any pair of goods (i and j) the absolute percentage difference between their consumption proportions ($k_i^* - k_j^*$) is less than the absolute percentage difference between the corresponding inverse shares of production ($x_i^* - x_j^*$). Then, the trade flow scaled by production (y_i^*) is a monotonic (though non-linear) transformation of the output ratios (z_i^*) across industries such that the ranking of y_i^* corresponds inversely to the ranking of z_i^*.*

PROOF:

Consider a discrete difference (Δ) between a typical pair of goods, i and j. For instance, define $\Delta y_{ij}^* = y_i^* - y_j^*$ $(i, j = 1, \ldots, n; i \neq j)$. Likewise, Δz_{ij}^*, Δk_{ij}^*, and Δx_{ij}^* can be defined in an analogous manner. Also define $\Delta(1 - y^*)_{ij} = (1 - y_i^*) - (1 - y_j^*) = -\Delta y_{ij}^*$. We wish to show that under the given condition of the proposition,

$$\text{(A1)} \quad \frac{\Delta y_{ij}^*}{\Delta z_{ij}^*} < 0 \qquad \text{for all } i \text{ and } j \; (i, j = 1, \ldots, n; i \neq j)$$

whence y_i^* is monotonically related to z_i^*. We first note that $\Delta z_{ij}^* = \Delta x_{ij}^*$. Hence, (15) holds whenever $(\Delta y_{ij}^*/\Delta x_{ij}^*) < 0$ for all i and j. This condition holds, in turn, whenever

$$\text{(A2)} \quad \frac{\Delta(1 - y^*)_{ij}}{\Delta x_{ij}^*} = -\frac{\Delta y_{ij}^*}{\Delta x_{ij}^*} > 0$$

Since trade outflow cannot exceed the level of production, and y_i^* is negative when there is an inflow, we have $y_i^* < 1$. Therefore, (A2) can be expressed in an equivalent form under the natural logarithmic transformation,

$$\text{(A3)} \quad \frac{\Delta \log (1 - y^*)_{ij}}{\Delta \log x_{ij}^*} \equiv \frac{\log (1 - y_i^*) - \log (1 - y_j^*)}{\log x_i^* - \log x_j^*} > 0$$

Applying the difference operator on the log transform of $(1 - y_i^*)$ in the revised equation (11), and after simplification, we obtain

$$\text{(A4)} \quad \frac{\Delta \log (1 - y^*)_{ij}}{\Delta \log x_{ij}^*} = \frac{\Delta \log k_{ij}^*}{\Delta \log x_{ij}^*} + 1$$

Since the first term on the right-hand side of (A4) may be positive or negative in general, we find that the inequality (A3), and hence (A1), will hold as long as $|\Delta \log k^*_{ij}| \leq |\Delta \log x^*_{ij}|$ in (A4), from which the proposition follows directly.

REFERENCES

B. Balassa, "An Empirical Demonstration of Classical Comparative Cost Theory," *Rev. Econ. Statist.*, Aug. 1963, *45*, 231–38.

R. E. Baldwin, "Determinants of the Commodity Structure of U.S. Trade," *Amer. Econ. Rev.*, Mar. 1971, *61*, 126–46.

J. Bhagwati, "The Pure Theory of International Trade: A Survey," *Econ. J.*, Mar. 1964, *74*, 1–84.

———, "The Heckscher-Ohlin Theorem in the Multi-Commodity Case," *J. Polit. Econ.*, Sept./Oct. 1972, *80*, 1052–55.

R. Bharadwaj, "Factor Proportions and the Structure of Indo–U.S. Trade," *Indian Econ. J.*, Oct. 1962, *10*, 105–16.

W. H. Branson and H. B. Junz, "Trends in U.S. Trade and Comparative Advantage," *Brookings Papers*, Washington 1971, *2*, 285–338.

——— **and N. Monoyios,** "Factor Inputs in U.S. Trade," *J. Int. Econ.*, May 1977, *7*, 111–31.

J. S. Chipman, "A Survey of the Theory of International Trade: Part 1, The Classical Theory," *Econometrica*, July 1965, *33*, 477–519.

Donald J. Daly, *International Comparisons of Prices and Output*, New York 1972.

H. Glejser, "Empirical Evidence on Comparative Cost Theory from the European Common Market Experience," *Euro. Econ. Rev.*, Nov. 1972, *3*, 247–58.

Herbert G. Grubel and Peter J. Lloyd, *Intra-Industry Trade*, New York 1975.

Gottfried Haberler, *The Theory of International Trade with Its Applications to Commercial Policy* (trans. from German), New York 1937.

J. Harkness, "Factor Abundance and Comparative Advantage," *Amer. Econ. Rev.*, Dec. 1978, *68*, 784–800.

——— **and J. F. Kyle,** "Factors Influencing United States Comparative Advantage," *J. Int. Econ.*, May 1975, *5*, 153–65.

E. F. Heckscher, "The Effect of Foreign Trade on the Distribution of Income" (trans. from Swedish), in Howard S. Ellis and Lloyd A. Metzler, eds., *Readings in the Theory of International Trade*, Philadelphia 1949, 272–300.

M. Hodd, "An Empirical Investigation of the Heckscher-Ohlin Theory," *Economica*, Feb. 1967, *34*, 20–29.

G. C. Hufbauer, "The Impact of National Characteristics and Technology on the Commodity Composition of Trade in Manufactured Goods," in Raymond Vernon, ed., *The Technology Factor in International Trade*, New York 1970, 145–231.

R. W. Jones, "Factor Proportions and the Heckscher-Ohlin Theorem," *Rev. Econ. Stud.*, No. 1, 1956, *24*, 1–10.

M. E. Kreinin, "The Theory of Comparative Cost—Further Empirical Evidence," *Econ. Int.*, Nov. 1969, *22*, 662–74.

E. E. Leamer, "The Commodity Composition of International Trade in Manufactures: An Empirical Analysis," *Oxford Econ. Papers*, Nov. 1974, *26*, 350–74.

W. W. Leontief, "Domestic Production and Foreign Trade: The American Capital Position Re-examined," *Proc. Amer. Phil. Society*, Sept. 1953, *97*, 332–49.

———, "Factor Proportions and the Structure of American Trade: Further Theoretical and Empirical Analysis," *Rev. Econ. Statist.*, Nov. 1956, *38*, 386–407.

G. D. A. MacDougall, "British and American Exports: A Study Suggested by the Theory of Comparative Costs, Part I," *Econ. J.*, Dec. 1951, *61*, 697–724.

——— **et al.,** "British and American Productivity, Prices, and Exports: An Addendum," *Oxford Econ. Papers*, Oct. 1962, *14*, 297–304.

L. W. McKenzie, "On Equilibrium in Graham's Model of World Trade and Other Competitive Systems," *Econometrica*, Apr. 1954, *22*, 147–61.

John Stuart Mill, *Principles of Political Economy with Some of Their Applications to Social Philosophy*, 3d ed., London 1852.

Bertil Ohlin, *Interregional and International*

Trade, Cambridge 1933.

Karen R. Polenske, *State Estimates of Technology, 1963*, Lexington 1974.

David Ricardo, *On the Principles of Political Economy, and Taxation*, London 1817.

Harry W. Richardson, *Regional Economics*, London 1969.

John M. Rodgers, *State Estimates of Interregional Commodity Trade, 1963*, Lexington 1973.

K. W. Roskamp, "Factor Proportions and Foreign Trade: The Case of West Germany," *Weltwirtsch. Arch.*, 1963, Part II, *91*, 319–26.

______ **and G. C. McMeekin,** "Factor Proportions, Human Capital and Foreign Trade: The Case of West Germany Reconsidered," *Quart. J. Econ.*, Feb. 1968, *82*, 152–60.

Sidney Siegel, *Nonparametric Statistics for the Behavioral Sciences*, New York 1956.

R. M. Stern, "British and American Productivity and Comparative Costs in International Trade," *Oxford Econ. Papers*, Oct. 1962, *14*, 275–96.

______, "Some Evidence on the Factor Content of West Germany's Foreign Trade," *J. Polit. Econ.*, Feb. 1976, *84*, 131–41.

M. Tatemoto and S. Ichimura, "Factor Proportions and Foreign Trade: The Case of Japan," *Rev. Econ. Statist.*, Nov. 1959, *41*, 442–46.

W. P. Travis, "Production, Trade, and Protection When There are Many Commodities and Two Factors," *Amer. Econ. Rev.*, Mar. 1972, *62*, 87–106.

D. F. Wahl, "Capital and Labour Requirements for Canada's Foreign Trade," *Can. J. Econ.*, Aug. 1961, *27*, 349–58.

T. M. Whitin, "Classical Theory, Graham's Theory, and Linear Programming in International Trade," *Quart. J. Econ.*, Nov. 1953, *67*, 520–44.

Incomplete Markets, Price Regulation, and Welfare

By H. M. POLEMARCHAKIS*

The normative appeal of the market mechanism rests on the equivalence between competitive equilibria and Pareto optimal allocations. This equivalence in turn depends on the existence of a complete system of markets; that is, on the possibility of trade in all commodities indexed by qualitative as well as temporal, locational, and contingent characteristics. On the other hand, problems of moral hazard and adverse selection, of incomplete and differential information, of transaction costs, and of aggregate, as opposed to individual, risk render the assumption of universality of markets empirically unjustified.

The question that follows is that of the characterization of situations where the allocations attained as price equilibria are dominated. In particular, I want to compare the price equilibrium allocations with the allocations attainable through the imposition of quantity constraints (i.e., rationing) in some of the markets. I shall demonstrate that in simple situations involving trade under uncertainty it may be preferable, in the absence of markets in contingent commodities, for prices to be regulated and for markets to be cleared through quantity rationing, as opposed to prices being allowed to fluctuate in response to the contingency realized.

The question of desirability of price stabilization has been a recurrent theme in economic theory. Leander Howell, Gertrud Lovasy, Walter Oi (1961, 1963, 1972), Clem Tisdell, and Frederick Waugh (1944a,b, 1966) posed the problem in a partial equilibrium context, while Paul Samuelson (1972a, b) raised the objection of feasibility in a general equilibrium framework and emphasized that interference with prices can only be beneficial in the absence of a complete system of markets.[1] In a different context, Martin Weitzman (1974) addressed the question of the optimal instruments of planning under uncertainty and compared quantity targets with profit maximization at appropriately chosen prices; he concluded that quantities dominate prices if the cost and benefit functions are sharply curved.

I address the problem in a general equilibrium framework. Further, since one of the alternatives considered involves trade at non-market-clearing prices, I employ the formalization developed by Jacques Dreze for market clearing through quantity constraints. I consider a consumption loan economy where differences in productivity (or, equivalently, initial endowments) among generations lead to price variability across time periods. As a result, when young, agents have to take an allocation decision under uncertainty; the prices they will have to face when old depend on the productivity of the generation then at the first period of its life. The question that arises is whether it is desirable to eliminate the future price uncertainty at the expense of constraining some agents below their desired level of transaction. The answer depends on the welfare criterion adopted and on the degree of risk aversion of the agents. I show that, under an expected utility criterion, the regime of price regulation and rationing dominates the regime of flexible prices, provided the agents are sufficiently risk averse concerning future consumption. On the other hand, individuals with the lowest productivity always prefer the flexible price equilibrium, while individuals whose productivity exceeds this minimum level prefer the regime of fixed prices if they happen to be sufficiently risk averse.

The intuition behind these results is clear. The higher the degree of risk aversion, the greater the losses that individuals suffer from their inability—due to the incompleteness of markets—to insure against future price fluctuations caused by fluctuations in productiv-

*Department of economics, Columbia University.

[1]See Giora Hanoch for a survey and synthesis of this literature.

ity. That insurance markets—or markets in contingent commodities—are absent is to be expected as long as it is aggregate fluctuations that affect the economy. Furthermore, one would guess that the higher the degree of diversity in the economy the more attractive regulation becomes. I demonstrate that this is indeed the case. Observe, finally, that the positive relation in my model between the desirability of regulation and the degree of diversity goes contrary to the results derived by Weitzman (1977). In his case the price mechanism succeeded to the extent that it managed to take into account differences among individuals in allocating a scarce resource; it failed according to some other criterion—equity for example. In my argument it is precisely differences among individuals (or generations) that, due to the incompleteness of markets, accentuate price fluctuation and, hence, render the price mechanism undesirable.

I. Variable Prices vs. Quantity Rationing

Consider an economy extending over time. During each time period, two generations are alive—one young and one old. A generation consists of one aggregate unit of identical agents, and there is no growth in population. Agents live for two periods, their endowment is their labor when young, and their intertemporal von Neumann-Morgenstern utility function has the form[2,3]

$$(1) \qquad u(l, c) = -l + \frac{1}{\alpha} c^{\alpha} \qquad \alpha < 0$$

where l is the labor supplied when young and c is the consumption when old. Young agents work so as to produce the consumption good according to the production function

$$(2) \qquad c = \gamma l \qquad \gamma > 0$$

which they sell to the old in exchange for fiat money. Old agents spend the money they earned when young so as to purchase the consumption good. The total stock of money is an exogenous constant[4] $M > 0$.

Uncertainty is introduced as follows: There are n types of generations, indexed by a subscript k, (i), $k = 1, \dots, n$, $(i = 1, \dots, n)$, and distinguished by the productivity of their labor; that is, by the parameter γ of the production function (2). There are no intragenerational differences. Consequently, the random variable $\tilde{\gamma}$ takes on the value γ_k with probability π_k. I shall assume that $\tilde{\gamma}$ is distributed independently over time.[5]

A price system for this economy takes the form $(p, w) = ((p_k, w_k), \; k = 1, \dots, n)$ where p_k and w_k are the money price of the consumption good and the money wage rate, respectively, when the young generation is of type $k = 1, \dots, n$. Given a price system, a young agent decides on how much labor to supply, $l^s_k(p, w)$, by solving the following maximization problem:

$$(3) \qquad \text{Max } -l + \frac{1}{\alpha} \sum_i \pi_i (w_k l / p_i)^{\alpha}$$

subject to $0 \leq l \leq 1$

Consequently, the supply of labor function is

$$(4) \qquad l^s_k(p, w) = \left[\sum_i \pi_i (w_k / p_i)^{\alpha}\right]^{(1/1-\alpha)}$$

A *flexible price equilibrium* can now be defined as a price system (p^*, w^*) such that both the good and the money markets clear, i.e.,

$$(5a) \qquad w^*_k \, l^s_k(p^*, w^*) = M \qquad k = 1, \dots, n$$

$$(5b) \qquad \gamma_k l^s_k(p^*, w^*) = M / p^*_k \qquad k = 1, \dots, n$$

The unique flexible price equilibrium can be computed as follows:

$$(6) \qquad p^*_k = M / (s\gamma_k) \qquad w^*_k = M / s \quad k = 1, \dots, n$$

[2]The linearity of the utility function with respect to labor is not important for the argument to follow. One can use $u(l, c) = -(1/\beta) l^{\beta} + (1/\alpha) c^{\alpha}$ and derive the same conclusions, only with more complicated algebra.

[3]It is necessary only that $\alpha < 1$. It is computationally convenient to assume that $\alpha < 0$.

[4]I am ignoring the possibility of active monetary policy. See Robert Lucas (1972) and the paper by Lawrence Weiss and myself for a discussion of the possibility of effective monetary policy in a similar model.

[5]The assumption of independence is not important. It is made so as to simplify definitions and computations, since it is not directly related to the points I want to demonstrate.

where s is defined in (7):

$$s = [\sum_i \pi_i \gamma_i^{\alpha}]^{(1/1-\alpha)} \tag{7}$$

At equilibrium, the amount of labor supplied and the level of expected utility derived are independent of k, and they take on the following values:

$$l^* = [\sum_i \pi_i \gamma_i^{\alpha}]^{(1/(1-\alpha))} \tag{8}$$

$$u^* = ((1-\alpha)/\alpha)[\sum_i \pi_i \gamma_i^{\alpha}]^{(1/(1-\alpha))} \tag{9}$$

Observe that, as expected, the level of the money stock M is neutral. Neither the equilibrium supply of labor nor the equilibrium consumption level are affected by changes in M. As a consequence, the flexible price equilibrium is essentially unique. This is important since I want to compare the utility level attained at a flexible price equilibrium with the alternative considered below.

I now want to look into the possible fixed-price equilibria for the economy; that is, equilibria characterized by $p_k = p_{k'}$, for any k and k'. Clearly, a quantity constraint must be imposed on some market. Furthermore, one would like to consider only "rational expectations equilibria": agents must not only know the prices that will prevail in the future; they must be able to indeed carry out, during the second period of their life, the plans they had formulated when they were young. Consequently, I shall analyze the case where only the supply of the consumption good is constrained.[6]

A fixed-price system has the form $(\bar{p}, w) = ((\bar{p}, w_k), k = 1, \ldots, n)$, where w_k is the money wage rate when the young generation is of type k, $k = 1, \ldots, n$, and p is the money price of the consumption good independently of k. In addition, an upper bound b is imposed on the amount of the consumption good that can be supplied in the market. Given a fixed-price system $(\bar{p}, w)$ and a constraint level b, a young agent decides on how much labor to supply, $\bar{l}_k^s(\bar{p}, w, b)$, by solving the following maximization problem:

$$\text{Max } -l + \frac{1}{\alpha}[w_k l/p]^{\alpha} \tag{10}$$

$$\text{subject to} \quad \gamma_k l \leq b \qquad 0 \leq l \leq 1$$

Consequently,

$$\bar{l}_k^s(\bar{p}, w, b) = \min[b/\gamma_k, (w_k/p)^{(\alpha/(1-\alpha))}] \tag{11}$$

A *fixed-price equilibrium* can now be defined as a fixed-price system $(\bar{p}^*, w^*)$, and an upper bound on the supply of the consumption good b^*, such that both the good and money markets clear, i.e.,

$$w_k^* \bar{l}_k^s(\bar{p}^*, w^*, b^*) = M \quad k = 1, \ldots, n \tag{12a}$$

$$\gamma_k \bar{l}_k^s(\bar{p}^*, w^*, b^*) = M/p^* \quad k = 1, \ldots, n \tag{12b}$$

A generation k is said to be constrained at an equilibrium $(\bar{p}^*, w^*, b^*)$ if and only if

$$(b^*/\gamma_k) < (w_k^*/\bar{p}^*)^{(\alpha/(1-\alpha))}$$

It is said to be unconstrained if and only if

$$(b^*/\gamma_k) \geq (w_k^*/\bar{p}^*)^{(\alpha/(1-\alpha))}$$

The fixed-price equilibrium is not unique. The following proposition, however, demonstrates that the problem of nonuniqueness can be circumvented.

PROPOSITION 1: (a) *There exists a continuum of constrained price equilibria indexed by the price* $\bar{p}$. (b) *There exists a unique constrained price equilibrium* $(\bar{p}^0, w^0, b^0)$ *where at least one generation is unconstrained.* (c) *The constrained price equilibrium* $(\bar{p}^0, w^0, b^0)$ *is preferred by every generation to any other constrained price equilibrium.*

PROOF:

Without loss of generality, we may suppose that

$$\gamma_n > \gamma_{n-1} > \ldots > \gamma_2 > \gamma_1 \tag{13}$$

The case where $\gamma_k = \gamma_{k'}$, for some $k \neq k'$ can be reduced to the one considered here by collapsing the generations k and k' into one

[6]It may appear as if an alternative way to the elimination of future price uncertainty exists. Namely, an upper bound is imposed on the demand of the consumption good by the old agents. This case is, however, equivalent to the one analyzed.

generation occurring with probability $\pi_k + \pi_{k'}$. Under this assumption the following is a fixed-price equilibrium:

$$\text{(14)} \qquad \bar{p}^0 = M(\gamma_1)^{(-1/(1-\alpha))}$$
$$w_k^0 = \bar{p}^0\gamma_k$$
$$b^0 = (\gamma_1)^{(1/(1-\alpha))}$$

The system of equations (12) is clearly satisfied since

$$\text{(15)} \quad \bar{l}_1^s = \min\left[\frac{(\gamma_1)^{(1/(1-\alpha))}}{\gamma_1}, \quad (\gamma_1)^{(\alpha/(1-\alpha))}\right] = (\gamma_1)^{(\alpha/(1-\alpha))}$$

$$\text{(16)} \quad \bar{l}_k^s = \min\left[\frac{(\gamma_1)^{(1/(1-\alpha))}}{\gamma_k}, \quad (\gamma_k)^{(\alpha/(1-\alpha))}\right] = \frac{(\gamma_1)^{(\alpha/(1-\alpha))}}{\gamma_k}, \qquad k = 2, \ldots, n$$

I now want to demonstrate that the fixed-price equilibrium ($\bar{p}^0, w^0, b^0$) is preferred by every generation to any other fixed-price equilibrium. Consider the utility function $u(l, c) = -l + (1/\alpha)c^\alpha$. Since, from the equilibrium conditions, $w_k = \bar{p}\gamma_k$, $k = 1, \ldots, n$, the utility function of generation k can be written as $u_k = -l + (1/\alpha)(\gamma_k l)^\alpha$. Observe that $(\partial^2 u_k/\partial l^2) < 0$ for all l. Furthermore, at the equilibrium($\bar{p}^0, w^0, b^0$) considered here, $(\partial u_1/\partial l) = 0$ while $(\partial u_k/\partial l) > 0$ for $k = 2, \ldots, n$. Consequently, generation 1 cannot improve its situation, while generation k ($k \geq 2$) can only improve its position by an increase in b above b^0. But this is not feasible. Given $b > b^0$, the labor supplied by generation 1 will not change, while the labor supplied by generation k, $k \geq 2$, will increase. Hence no fixed market-clearing prices will exist. Equivalently, the fixed-price equilibrium defined in (14) Pareto dominates any other fixed-price equilibrium. For any $b^* < b^0$ setting $\bar{p}^* = M/b^*$ and $\bar{w}_k = \bar{p}^*\gamma_k$, $k = 1, \ldots, n$, we see that $\bar{l}_k^s = b^*/\gamma_k$, $k = 1, \ldots, n$, and hence an equilibrium ensues. Consequently, there exists a continuum of equilibria indexed by b^* or $\bar{p}^* = M/b^*$. Finally, it is clear from (15) that at the fixed-price equilibrium ($\bar{p}^0, w^0, b^0$) generation 1 is not constrained and from (16) that all others are. At any other fixed-price equilibrium, $b^* < b^0$ and hence all generations are constrained.

We now have a well-defined fixed-price equilibrium to compare, from a welfare point of view, with the flexible price equilibrium. It is not evident, however, what the appropriate welfare criterion is. In particular, is one to look for dominance in the expected utility (or *ex ante*) sense, or for dominance in the unanimity (or *ex post*) sense? I believe that expected utility is the appropriate criterion in situations involving risk. *Ex post* criteria may lead to absurd conclusions. On the other hand, a case for the unanimity criterion can be made in the context of the model at hand since every agent knows without ambiguity which type he belongs to. I shall consequently consider both criteria.

Before stating the results, let me give an intuitive argument. Under flexible prices, the supply of labor depends only on agents' expectations about the future, and hence, is independent of the type of generation they themselves belong to. This is clear from (8) where l^* is independent of k. But if labor supplied is independent of k, future consumption is a random variable whose distribution is nondegenerate as long as the productivity parameter has a nonzero variance. Furthermore, since people are risk averse concerning future consumption ($\alpha < 1$), this uncertainty entails a loss in utility and consequently it may be better to stabilize future consumption at the expense of variability of labor supplied. If the utility function were not linear in labor but had the form $u(l, c) = -(1/\beta)l^\beta + (1/\alpha)c^\alpha$, the profitability of stabilization would depend not on α alone but on $(\alpha - \beta)$. Finally, it is clear that the only generation which is certain to lose from stabilization is the generation with the lowest productivity: It can no longer take advantage of the higher productivity of its descendents. I shall now formalize these arguments in the two propositions to follow.

PROPOSITION 2: (a) *For any level of risk aversion (i.e., for any* $\alpha < 1$), *an unconstrained generation prefers the flexible price equilibrium to the fixed-price equilibrium* ($\bar{p}^0, w^0, b^0$). (b) *For a high level of risk*

aversion (i.e., for $\alpha << 0$),[7] *a constrained generation prefers the fixed-price equilibrium ($\bar{p}^0, w^0, b^0$) to the flexible price equilibrium.*

PROOF:

From (15), (16), and (1), it can be computed that the level of utility, $\bar{u}_k$, attained by generation k at the fixed-price equilibrium ($\bar{p}^0, w^0, b^0$) is as follows:

$$(17)\quad \bar{u}_1 = ((1-\alpha)/\alpha)(\gamma_1)^{(\alpha/(1-\alpha))}$$

$$(18)\quad \bar{u}_k = ((\gamma_k - \alpha\gamma_1)/\alpha\gamma_k)(\gamma_1)^{(\alpha/(1-\alpha))},\quad k = 2, \ldots, n$$

From (9), (17) and (18) we get that

$$(19)\quad u^*/\bar{u}_1 = [\pi_1 + \sum_{k=2}^{n} \pi_k(\gamma_k/\gamma_1)^\alpha]^{(1/(1-\alpha))}$$

$$(20)\quad u^*/\bar{u}_k = ((\gamma_k - \alpha\gamma_k)/(\gamma_k - \alpha\gamma_1)) \cdot [\pi_1 + \sum_{k=2}^{n} \pi_k(\gamma_k/\gamma_1)^\alpha]^{(1/(1-\alpha))},\quad k = 2, \ldots, n$$

Since u^* and $\bar{u}_k$, $k = 1, \ldots, n$, are always negative, $(u^* > \bar{u}_1)$ if and only if $((u^*/\bar{u}_1) < 1)$, and $(\bar{u}^* < \bar{u}_k)$ if and only if $((u^*/\bar{u}_k) > 1)$, $k = 2, \ldots, n$. From (19), $((u^*/\bar{u}_1) < 1)$ if $((\gamma_k/\gamma_1)^\alpha < 1)$ for $k = 2, \ldots, n$, which holds for all $\alpha < 0$ since, by assumption, $((\gamma_k/\gamma_1) > 1)$ for $k = 2, \ldots, n$. Consequently, the generation which is unconstrained at the fixed-price equilibrium prefers the flexible price equilibrium independently of its degree of risk aversion. On the other hand, as α tends to $-\infty$, the term $((\gamma_k - \alpha\gamma_k)/(\gamma_k - \alpha\gamma_1))$ of (20) tends to $(\gamma_k/\gamma_1) > 1$, while the term

$$[\pi_1 + \sum_{k=2}^{n} \pi_k(\gamma_k/\gamma_1)^\alpha]^{(1/(1-\alpha))}$$

tends to 1 for all $k = 2, \ldots, n$. Consequently, there exists $\bar{\alpha}_k < 0$ such that $((u^*/\bar{u}_k) > 1)$ for $\alpha < \bar{\alpha}_k$. A generation which is constrained at the fixed-price equilibrium prefers that to the flexible price equilibrium, provided it is highly risk averse.

[7]Relative risk aversion is defined as $-u''y/u'$. Hence, if $u(l, c) = -l + (1/\alpha)c^\alpha$, risk aversion concerning future consumption is $(1-\alpha)$. Consequently, as α tends to $-\infty$, relative risk aversion tends to $+\infty$.

PROPOSITION 3: *If the risk aversion of agents is high (i.e., for $\alpha << 0$), the regime of fixed prices and rationing dominates the regime of flexible prices.*

PROOF:

Given the results of the previous propositions, the argument is now immediate. As α tends to $-\infty$, $(u^*/\bar{u}_1)$ tends to 1, while $(u^*/\bar{u}_k)$ tends to (γ_k/γ_1) which is greater than 1, for $k = 2, \ldots, n$. Equivalently, there exists $\bar{\alpha} < 0$ such that if $\alpha < \bar{\alpha}$, $(\Sigma_{k=1}^{n} \pi_k\bar{u}_k) > u^*$. Consequently, for $\alpha < \bar{\alpha}$, the level of expected utility attained under a regime of fixed prices and rationing exceed the level of expected utility attained under a regime of flexible prices.

II. Diversity and the Desirability of Regulation

The standard argument in favor of the price mechanism as opposed to quantity rationing in the allocation of resources goes somewhat as follows: The price system allocates a scarce resource differentially among agents giving more of the commodity to those who demand it most. Rationing on the other hand cannot take into account differences among individuals. This argument is, of course, misleading. It presupposes that it is easier to compute equilibrium prices than equilibrium quantities, even though, as is well known, the information required is the same for both. Differently put, it contrasts not the price mechanism with the rationing mechanism but a sophisticated price mechanism with a naive rationing mechanism. All that can be justifiably argued on a priori grounds is that the price system is not inferior to a rationing mechanism.

Weitzman (1977) looks into the choice between prices and quantity rationing from a different perspective. He draws a distinction unfamiliar to standard consumer theory between "needing" a commodity and "being willing to pay a high price" for it. He argues that the latter can be the outcome either of need or of high nominal income. If the second is the case, and if society has decided to try to approximate the allocation that would result from the competitive mechanism under a

uniform distribution of income, the price system may be inappropriate. In more general terms, if society has chosen a particular allocation as its objective, even if the latter is Pareto efficient, the price system cannot always be relied upon to attain it. On the other hand, given this objective function, the higher the variation in the distribution of needs in the population the more attractive does the price mechanism become. Peter Stan deals with the same problem in a more economic framework. He postulates a market with costly transactions and compares it with a rationing mechanism. As expected, he shows that the more diverse the preferences and endowments of agents, the greater the relative merit of the price mechanism.

The model presented in the previous section offers an alternative framework within which to raise the same question: How does the desirability of the price system depend on the degree of diversity in the economy? Observe now that the intuition behind my argument is the reverse of that behind the argument of Weitzman and Stan. In their case, prices are successful in so far as they take diversity among agents into account, but they fail according to some other criterion—for example, equity or transaction costs. As a consequence, the desirability of prices depends positively on the degree of diversity in the economy. It is precisely, however, the diversity among agents (generations) that renders—in my model—the price mechanism undesirable. The more diverse generations in their productivity, the higher the fluctuations in prices needed to clear the markets; and the more pronounced the fluctuations, the higher the welfare losses due to the absence of markets in contingent commodities. As a consequence, one expects that the larger the variation in productivity among generations, the more desirable price regulation becomes. I shall conclude by formalizing this claim.

Two problems arise in giving a precise statement of the positive relation between diversity and the desirability of regulation. First, how is diversity to be measured? As A. B. Atkinson has argued, the measure that is appropriate depends on the utility function of individuals and the production technologies of firms. The only structure that can be imposed without loss of generality is that there exists an increasing and concave function $U(y)$, and that distributions $f(y)$ are ranked according to the diversity measure

$$D = \int U(y) f(y) dy \tag{21}$$

I shall say a measure of diversity is well defined if it is of the form (21) with the function $U(y)$ increasing and concave.

The second problem that arises is that of the distinction between a difference in the level of diversity between two economies and a difference in the general level of productivity.[8] An economy ϵ in my framework is characterized by the degree of risk aversion of the agents, α, the possible values of the productivity parameter γ_k, $k = 1, \ldots, n$, and by the probability of occurrence of each type of generation π_k, $k = 1, \ldots, n$. To compare the effects of different degrees of diversity between two economies, we must hold constant the average level of productivity—measured in utility terms—as well as the degree of productivity of the least productive generation. Consequently, I shall say that two economies $\epsilon = (\alpha, (\gamma_k, \pi_k), k = 1, \ldots, n)$ and $\epsilon' = (\alpha', (\gamma'_k, \pi'_k), k = 1, \ldots, n)$ are similar if and only if

$$\alpha = \alpha' \tag{22a}$$

$$\min_k (\gamma_k) = \min_k (\gamma'_k) \tag{22b}$$

$$u^*(\epsilon) = u^*(\epsilon') \tag{22c}$$

The following proposition relates the degree of diversity in an economy with the desirability of price regulation.

PROPOSITION 4: *There exists a well-defined measure of diversity, D, with the following properties:* (a) *Given two economies, ϵ and ϵ', which are similar, if the degree of diversity in ϵ, $D(\epsilon)$, is at least as high as the degree of diversity in ϵ', $D(\epsilon')$, dominance, in the expected utility sense, of the fixed-price mechanism over the flexible price mechanism in ϵ' implies dominance of the fixed-price mechanism in ϵ as well.* (b) *Given*

[8]The same holds for the argument of Weitzman (1977).

two economies, ϵ and ϵ', with the same minimum productivity level, if the degree of diversity in ϵ, $D(\epsilon)$, exceeds the degree of diversity in ϵ', $D(\epsilon')$, there exists a finite constant $\bar{\alpha}$ such that, if $\alpha < \bar{\alpha}$ and $\alpha' < \bar{\alpha}$, the gains from price regulation in the economy ϵ exceed the gains from price regulation in the economy ϵ'.

PROOF:

Consider the measure of diversity defined by

$$(23) \qquad D(f(y)) = \int -(1/y)f(y)dy$$

Since the function $U(y) = -1/y$ is increasing and concave for $y > 0$, we may define the degree of diversity in an economy ϵ by

$$(24) \qquad D(\epsilon) = -\sum_{k=1}^{n} (\pi_k/\gamma_k)$$

The level of expected utility attained at economy ϵ under a regime of fixed prices $\bar{u}(\epsilon)$ can be written as

$$(25) \qquad \bar{u}(\epsilon) = (\gamma_1)^{(\alpha/(1-\alpha))}[1/\alpha + \gamma_1 D(\epsilon)]$$

while the gains from price regulation, $-\bar{u}(\epsilon)/u^*(\epsilon)$, can be written as

$$(26) \qquad -\bar{u}(\epsilon)/u^*(\epsilon) = \left(\frac{\alpha}{\alpha - 1}\right) \gamma_1^{(\alpha/(1-\alpha))} \cdot [1/\alpha + \gamma_1 D(\epsilon)] \left[\sum_{k=1}^{n} \pi_k \gamma_k^{\alpha}\right]^{1/(\alpha-1)}$$

From (26), as $\alpha \rightarrow -\infty$, $-\bar{u}(\epsilon)/u^*(\epsilon)$ tends to $[D(\epsilon)/\gamma_1]$ and hence part (b) of the proposition follows. Part (a) is clear from (25).

III. Conclusion

Two major objections can be raised against the arguments presented in this paper. First, it may be argued that the model considered is so simplistic that any extrapolation to "real" economic situations is suspect. Second, it may be argued that the implementation of the fixed-price equilibrium requires a larger amount of information than that of the flexible price equilibrium. As far as the first objection is concerned, it is clear that one can use general utility and production functions and derive the same results by taking Taylor's approximations. What can not be dispensed with is the highly aggregated structure of the economy. It is not even clear how the problem is to be posed in a more disaggregated model. Concerning the informational requirements of regulation, it suffices to point out that knowledge of the flexible price equilibrium is sufficient for the computation of the fixed-price equilibrium, as well as of the prevailing degree of risk aversion.

REFERENCES

A. B. Atkinson, "On the Measurement of Inequality," *J. Econ. Theory*, Sept. 1970, *2*, 244–63.

J. Dreze, "Existence of an Exchange Equilibrium under Price Rigidities," *Int. Econ. Rev.*, June 1975, *16*, 301–20.

G. Hanoch, "Desirability of Price Stabilization or Destabilization," disc. paper no. 351, Harvard Instit. Econ. Res. 1974.

L. D. Howell, "Does the Consumer Gain from Price Instability?," *Quart. J. Econ.*, Feb. 1945, *59*, 287–95.

G. Lovasy, "Further Comment," *Quart. J. Econ.*, Feb. 1945, *59*, 296–301.

R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

W. Oi, "The Desirability of Price Instability Under Perfect Competition," *Econometrica*, Jan. 1961, *29*, 58–64.

———, "Uncertainty, Instability, Expected Profit: Rejoinder," *Econometrica*, Jan./Apr. 1963, *31*, 248.

———, "The Consumer Does Benefit from Feasible Price Stability: A Comment," *Quart. J. Econ.*, Aug. 1972, *86*, 494–98.

H. M. Polemarchakis and L. Weiss, "On the Desirability of a Totally Random Monetary Policy," *J. Econ. Theory*, Aug. 1977, *15*, 345–50.

P. A. Samuelson, (1972a) "The Consumer Does Benefit from Feasible Price Stability," *Quart. J. Econ.*, Aug. 1972, *86*, 476–93.

———, (1972b) "The Consumer Does Benefit from Feasible Price Stability: Rejoinder," *Quart. J. Econ.*, Aug. 1972, *86*, 500–03.

P. J. E. Stan, "Welfare Economics When Transactions are Costly: Criteria for Choice Among Alternative Institutions,"

mimeo., Harvard Univ. 1978.

C. **Tisdell**, "Uncertainty, Instability, Expected Profits," *Econometrica*, Jan./Apr. 1963, *31*, 243–47.

F. V. **Waugh**, (1944a) "Does the Consumer Benefit from Price Instability?," *Quart. J. Econ.*, Aug. 1944, *58*, 602–14.

______, (1944b) "Reply," *Quart. J. Econ.*, Feb. 1945, *59*, 296–301.

______, "Consumer Aspects of Price Instability," *Econometrica*, Apr. 1966, *34*, 504–08.

M. L. **Weitzman**, "Prices vs. Quantities," *Rev. Econ. Stud.*, Oct. 1974, *41*, 477–92.

______, "Is the Price System or Rationing More Effective in Getting a Commodity to Those who Need it Most?," *Bell J. Econ.*, Autumn 1977, *8*, 517–24.

The Measurement of Inequality: Comment

By KENNETH L. WERTZ*

There appears to be an inconsistency in Morton Paglin's recent article in this *Review*. The effect of that inconsistency is to exaggerate the difference between the traditional Gini coefficient and a Gini coefficient which is adjusted for the average age-earnings relationship.

Paglin observes that the Gini coefficient is often used in normative evaluations of the income distribution. What begins as a measure of inequality is finally treated, with or without intervening statements of qualification, as if it were a measure of inequity. This being so we must ask whether all deviations from the mean income of the population—the inequalities that give magnitude to the Gini coefficent, G—are unjustifiable.

For Paglin the answer is no, and on this point one suspects that he belongs to a comfortable majority. The question is what to do. Paglin proposes that use be made of the average age-income relationship. Perfect equality or equity exists when all families at the same stage in their life cycle have the same annual income. The deviation of any one family's income from the mean income for its age cohort is taken to be an unjustifiable deviation that gives magnitude to some adjusted Gini coefficient. Since the actual age-income profile has a definite hump, it is expected that the adjusted Gini coefficient will be less than G.

My points will be that Paglin, in giving mathematical expression to his normative position, has made a structural change in the Gini formula which does not fully conform with the logic of the Gini measure, and that in consequence the magnitude of inequality which he has measured is determinately too small.

Suppose that the income scale and the age scale are partitioned into a finite number of segments. Each family belongs to one of the income ranges and to one of the age (of head) ranges so defined. Let n_{ij} represent the number of families that have an annual income in income range i and whose family heads are in age range j. They will be said to be "in cell (i, j)." The income level of every family in cell (i, j) is denoted by y_{ij}.[1] The mean income level for all families in age range j is denoted m_j, and the grand mean is m. There are N families in the population.

The Gini coefficient,

$$(1)\quad G = \sum_j \sum_i n_{ij} \sum_r \sum_k n_{kr} \cdot |y_{ij} - y_{kr}| / 2mN(N-1)$$

viewed as a measure of inequity in the distribution of income, supposes that every family should have an income of m. Select arbitrarily a family in cell (i, j) as the referent family. Its income is compared with that of another family in, say, cell (k, r). One records in absolute value the difference $(|y_{ij} - y_{kr} - 0|)$ between (a) the actual differential $(y_{ij} - y_{kr})$ in their incomes and (b) the differential $(m - m = 0)$ that would obtain under perfect equity. After pairwise comparisons between the referent family and all other families have been made, and after all families have served as the referent family, the differences are cumulated to form the numerator of (1). Dividing by $2m$ times the total number of such comparisons standardizes the range of G to the interval [0, 1].

Paglin advances the idea that the equitable income of every family in age range j is m_j. Following the logic of the Gini coefficient step for step—except that the differential in equitable incomes is now $m_j - m_r$—we would be led to an adjusted Gini coefficient,

*Associate professor, University of North Carolina-Chapel Hill.

[1]Ideally the age and income ranges would be made as finely as the significance of the data allows. If the ranges were any wider there would be intracell variations in incomes and ages. Use of the formulas (1)–(3) below would then require that those variations be suppressed.

$$(2)\quad ADJG = \sum_j \sum_i n_{ij} \sum_r \sum_k n_{kr} |(y_{ij} - y_{kr}) - (m_j - m_r)| / 2mN(N-1)$$

Paglin, however, does not arrive at $ADJG$. Instead he formulates an age Gini (i.e., the value a traditional Gini coefficient would take if every family had an income equal to its respective m_j) and subtracts it from G to obtain the Paglin Gini (PG) (p. 601, fn. 3). In present notation,

$$(3)\quad PG = \sum_j \sum_i n_{ij} \sum_r \sum_k n_{kr}(|y_{ij} - y_{kr}| - |(m_j - m_r|)/2mN(N-1)$$

The problem with PG is that it does not consistently implement the new normative view that motivated the adjustment in the Gini coefficient. A numerical example illustrates. Abel's income is \$20,000 and the average in his age range is \$19,000. Baker's income is \$15,000 and the average in his age range is \$10,000. Their incomes should differ under the new normative view by \$9,000, but they actually differ by just \$5,000. The Abel/Baker comparison, with Abel as the referent, therefore contributes \$4,000 to the numerator of $ADJG$, as intended. However it contributes $-$\$4,000 to the numerator of PG, the rationale for which is difficult to perceive. This anomalous result is repeated in the Baker/Abel comparison when Baker becomes the referent.

The mathematical divergence of (3) from (2) comes in the placement of absolute value signs. Let $A = y_{ij} - y_{kr}$ and $B = m_j - m_r$. By a variant of the triangle inequality, $|A - B| \geq |A| - |B|$. Inequality holds when

$$(4)\quad B > 0 \text{ and } B > A$$

$$\text{or} \quad B < 0 \text{ and } B < A$$

These are inexact conditions, covering six of the thirteen[2] logically distinct relations between A, B, and zero. More decisively, because the average age-earnings relationship has curvature and because age-cohort income distributions overlap, conditions (4) must in fact be satisfied for many of the comparisons which involve families in different age ranges.[3] Thus $ADJG > PG$: Paglin's formulation produces values of an adjusted Gini coefficient that are significantly too low, given the normative view that prompted the adjustment. It follows that the difference between G and PG when computed for a given year overstates the effect of accounting for the average age-earnings relationship within the logic of a Gini coefficient. Nevertheless it may still be the case that the consistently adjusted Gini coefficient would show, as does the Paglin Gini, a larger percentage decline over time than the unadjusted Gini coefficient.

It is important to realize that a comparison between $ADJG$ and G is a comparison of magnitudes of inequity that are to be associated with a given reality (the income and age distributions that exist in a particular period). Although the two indicators bring different norms to the distributions, they both compute the magnitude of inequity as one-half the average absolute value of inequitable differentials between all pairs of families relative to overall mean income. Any difference in their values shows by how much the magnitude of inequity is affected when we switch norms but keep the same type of measure.

However $ADJG$, unlike G, does not indicate the degree of inequity to be associated with a given reality. The coefficient G indicates degree because it is the ratio of two magnitudes of inequity, one (G) associated with the actual distribution and the other (unity) associated with the hypothetical distribution wherein all income is concentreated with just one family. In other words the maximal value of the traditional Gini under full concentration is unity, so that "what is" is normalized by "what at worst could have been." In contrast, the maximal value of $ADJG$ under full concentration is essentially 2. Let J denote the age cohort of the one family that hypothetically has all

[2]Namely, $A \gtrless B$ with $A, B > 0$ or $A, B < 0$; $A > 0$ with $B \leq 0$; $A = 0$ with $B \gtrless 0$; and $A < 0$ with $B \geq 0$.

[3]However the overlapping of age-cohort income distributions is not necessary for (4) to obtain.

income and define $n_j = \Sigma_i n_{ij}$. It can be shown that

(5) $$ADJG = (2N + 1 - n_j - (2N/n_j))/(N - 1)$$

in that case. Expression (5) attains a maximum when $n_j = \sqrt{2N}$, namely

$$\max ADJG = (2N + 1 - 2\sqrt{2N})/(N - 1)$$

and this quantity approaches 2 as N becomes larger.[4] Hence G should be compared with $ADJG/2$ for large values of N in the event that degrees of inequity are to be compared for the same distributions. Of course the deflator is inconsequential when percentage changes in degree across time periods are being calculated and compared with percentage changes in the traditional Gini.

[4]Max $ADJG$ exceeds 1 99 for N greater than 8×10^4.

REFERENCE

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

The Measurement of Inequality: Reply

By MORTON PAGLIN*

Kenneth Wertz questions the method by which I remove life cycle effects from the Gini coefficient. He argues that my *PG* formula "does not fully conform with the logic of the Gini measure, and . . . is determinatively too small" (p. 670). An alternative Gini (*ADJG*) is proposed which implements the normative view of the new equality standard and follows "the logic of the Gini coefficient step for step" (p. 670). I will show that 1) Wertz's argument derives no support by appealing to the logic of the Gini, 2) *ADJG* (as well as the average of the cohort Ginis) has a built-in upward bias making it unsuitable as a measure of long-term inequality or of perceived inequity, and 3) the fact that *PG* is based on the expression $\Sigma\ |A|\ -\ \Sigma|B|$ rather than $\Sigma\ |A\ -\ B|$ endows it with a useful property which makes *PG* a better measure of long-term inequality than alternative Gini formulas; this is confirmed by longitudinal income studies.

The Gini coefficient followed the Lorenz curve both historically and conceptually. In recognition of its antecedent, the Gini is appropriately defined as the area between the idealized *L* curve of equal distribution and the *L* curve of the actual distribution, divided by the area under the curve of equal distribution. Noteworthy is the conceptual distinctiveness of the two cumulative income functions—the actual and the ideal—which delineate this area of inequality. On this basis, the *PG* formula which allows independent consideration of the actual income differences (resulting in the *L* curve) and the idealized differences (*P* curve) seems closer to Gini's intent than Wertz's formula which merges the two by modifying the actual incomes by the idealized differences when pairing incomes; furthermore, as noted below, the *L* curve describing the lower boundary of *ADJG* may fall outside the traditional axes of the Lorenz diagram. In a graphic representation of *ADJG*, the *L* curve of actual incomes disappears since Wertz's coefficient is based on a modified income distribution from which cohort mean income differences have been removed. This can be seen by analysis of Wertz's equation (2). The numerator of *ADJG* employs the expression $|\ (y_{ij} - y_{kr}) - (m_j - m_r)\ |$ which is equal to $|\ (y_{ij} - m_j) - (y_{kr} - m_r)\ |$. Thus actual incomes are first expressed as deviations from their cohort means before the pairing process. In order to show *ADJG* on a Lorenz diagram it is necessary to derive the modified incomes (y') which underlie or correspond to the adjusted differences used by Wertz. For families in the jth cohort, the transformed incomes are

$$(1) \qquad y'_{ij} = m + (y_{ij} - m_j)$$

where m is the overall mean and m_j the cohort mean. Note that the $(y_{ij} - m_j)$ values can be added to any constant without affecting the numerator of *ADJG* but m is the only constant which leaves the denominator unchanged as well; hence, by using m, the means of the actual and the derived distributions are kept the same. The *L* curve of all y' values together with the 45° line define the area of inequality of the *ADJG* coefficient, a measure supposedly free of life cycle effects. The *PG* on the other hand keeps the *L* curve of actual incomes and compares it with a new equality standard which allows income differences between, but not within, age cohorts. While the area of *PG* always falls within the conventional boundaries of the Lorenz diagram, *ADJG* may dip below the base line into the negative income quadrant.[1]

*Portland State University

[1]Wertz states that *ADJG* has a maximum value of 2 rather than 1. On a Lorenz diagram this means that the area of inequality bounded by the *L* curve can be greater than the traditional area of maximum concentration. How is this possible? Wertz's adjusted incomes shown in my equation (1) indicate the reason. The y' values can be negative for families with zero or low incomes in cohorts where $m_j > m$. This follows since $(y_{ij} - m_j)$ for these families will be negative and of greater absolute value than m. Hence the *L* curve of y' values will dip below the

Tradition aside, adjusting the income curve is acceptable if the adjustments roughly correct for stage-of-life cycle differences in the annual income data. Wertz's adjustments do not. He removes the mean income differences from the cohort distributions—in effect lining up the cohorts on a flat age-income profile at the mean level—and calculates the residual inequality on the assumption that within-cohort distributions would be unaffected by eliminating the curved life cycle pattern. But if we follow the correct procedure of annualizing or flattening each individual life cycle curve, the cohort distributions will undergo a marked reduction in dispersion, except in the special case where all individual life cycle curves are parallel.[2] This indicates the upward bias in *ADJG*; a more revealing method is described later.

One way of using an equality standard for equity purposes is to compare "what is" with "what would be" under the idealized standard. The *PG* formula as I have shown (1977, p. 521) can be viewed in terms of an income comparison game in which each person compares his score (expected gain) in the actual distribution with his expected gain under the idealized standard. The difference between these two scores is a reasonable way of assessing current inequality. Hence on a micro level *PG* has a clear meaning and conforms to the logic of the Gini.[3] In Wertz's example, Baker (let us say age 30) earns $15,000 and Abel (age 40) earns $20,000. Under the new standard Baker's income would be $10,000 and Abel's would be $19,000, these being the mean incomes of their cohorts. Although most income differences would be narrowed or eliminated under the *P*-equality standard, some would move the other way, and every change has to be recorded. The Baker/Abel gap would widen from $5,000 to $9,000; hence Baker in the idealized distribution would be $4,000 poorer vis-à-vis Abel than he is currently. This explains the negative $4,000 contributed to the numerator of *PG* by this pairing since *PG* represents the difference between current inequality and the limited age-related inequality permitted under the standard. (In Wertz's implementation, Baker is supposed to feel closer to Abel after the income difference between them has widened.)

At this point, Wertz might argue that subjective perceptions of income differences are not central to the problem of measuring long-term inequality; and "objectively" according to the new standard Baker's $15,000 is really $4,000 greater than Abel's $20,000 once we eliminate the $9,000 difference in the mean incomes of their cohorts. Baker is $5,000 above his cohort mean, but Abel is only $1,000 above the mean of his cohort. Therefore Baker, no matter what he may

---

zero income level before swinging upward to the right-hand corner. Under the conditions of maximum concentration specified by Wertz, the *L* curve will drop sharply almost the full length of the negative income scale, then move at a 45° angle to the zero income line, and finally straight up to the top right corner, the resulting area of inequality will approximate a parallelogram twice the size of the usual triangle. Hence Wertz suggests using *ADJG*/2 when comparing his coefficient with *G*. This would not generally be advisable since for typical lognormal income distributions, *ADJG* values are unaffected by the higher theoretical limit and fall below *G*; for 1972 family incomes, *ADJG* is .324, closely matched by the weighted average of the cohort Ginis, .334. This is not surprising since both eliminate the effect of differences in cohort income levels, one by shifting cohort distributions to a common base, the other by calculating restricted within-cohort Ginis on different cohort means rather than a common mean. The *ADJG* has the advantage of allowing unrestricted between-cohort income pairing.

[2]For example, a teamster and a college professor may differ by only a small amount in annualized lifetime incomes, but they will typically show wider actual differences in each age cohort because of different life cycle curves.

[3]The graphic definition of the Gini which emphasizes the separate identity of the actual and the ideal distributions implies that each person views inequality by comparing the set of actual income differences with the set of ideal differences which would prevail under the equality standard. Referring to Wertz's argument, it may be said that *PG*, based on this conceptual separation, namely

$$\sum_j \sum_i n_{ij} \sum_r \sum_k n_{kr} |y_{ij} - y_{kr}| - \sum_j n_j \sum_j n_j \sum_r n_r |m_j - m_r|$$

more easily conforms to the logic of the Gini than *ADJG* which is based on

$$\sum |(y_{ij} - y_{kr}) - (m_j - m_r)|$$

In the latter expression, separate consideration of the actual and the idealized distributions is not possible. At best, such reasoning is only peripheral; the two formulas must be explained on their own terms.

perceive, should be considered $4,000 richer than Abel.

What is wrong with this analysis? Adjusting incomes across age cohorts in order to estimate long-term inequality (or inequity) would be a reasonable method provided that we had data on individual life cycle income curves; but with limited cross-section data, Wertz's *ADJG* estimates will be grossly biased upward and inferior to my *PG*. This stems from his implicit assumption that individual life cycle curves will all be roughly parallel. For example, Wertz's conclusion that 30-year old Baker (with $15,000) is really $4,000 richer than 40-year old Abel (income $20,000) assumes that when Baker reaches 40 he will earn $24,000 or occupy the same relative income slot ($5,000 above the mean) which he now occupies in the age 30 cohort. If life cycle curves cross, the errors generated by this assumption become apparent. Suppose Baker is a factory worker with a relatively flat life cycle income; he knows that at age 40 he will earn less than Abel and not $4,000 more. Why then should he consider himself richer than Abel—or why objectively should we? Similarly, in the Abel/Baker pairing, according to Wertz, Abel is supposed to consider himself $4,000 poorer than the younger Baker; again, this assumes that at age 30 Abel had an income of exactly $11,000 or that he occupied the same relative income slot ($1,000 above the cohort mean) which now marks his place in the age 40 cohort. The upward bias in Wertz's adjusted Gini results from the use of a single point in a person's life cycle to infer the shape of his whole life cycle income; this is done when making adjusted income pairings by implicitly specifying past and future income points with reference to other cohort means. Since we have income pairings for all persons, and across all cohorts, we are implicitly extrapolating parallel life cycle income curves for all.[4] What is the quantitative significance of this assumption? Recent longitudinal income studies provide some pertinent answers. Before reviewing these it will be helpful to sketch briefly the way inferences about lifetime incomes are made.

Abstracting from economic growth, we infer life cycle incomes from cross-section data by assuming that as a cohort ages it will occupy the income chairs now held by the older cohorts. But we also need information about the expected degree of intracohort mobility as a cohort ages. Zero mobility is indicated by perfect rank-order correlation of incomes across all years: a person occupying the *n*th income rank in his cohort at 30 would always occupy the *n*th income rank as the cohort moved through the life cycle. Individual life cycle curves would never cross and lifetime income inequality would be at a maximum, given the cohort income distributions. Under these conditions Wertz's *ADJG* or its proxy, the average of the cohort Ginis weighted by income shares, would closely measure lifetime inequality.[5] Once we allow some intracohort mobility, lifetime inequality typically will drop, and if the rank-order transition matrix exhibits very high mobility, lifetime incomes will converge toward equality.

Three recent studies show the importance of intracohort mobility for inequality. Bradley Schiller found that rank-order changes occurred across the entire range of the income scale and mobility could be viewed as a pervasive dynamic characteristic of our distribution. Donald Parsons' longitudinal income and autocorrelation study bears more directly on the issues raised here. Using male earnings from the National Longitudinal Surveys, Parsons found that: "The distribution of lifetime human wealth depends not only on the distribution of annual earnings but also on the consistency with which individuals maintain their economic position in the distribution from year to year" (p. 551). His results "indicate that the actual standard deviation

[4]By contrast, *PG* does not attempt to adjust actual incomes across cohorts to find the age-equivalent incomes since to do this properly requires knowledge of each individual's life cycle income; instead *PG* simply compares the average of the actual differences between two cohorts with the average of the ideal differences. This also seems crude, but with limited cross-section data it is less biased and provides better estimates of long-term inequality than does *ADJG*.

[5]For *CPS* income distributions, *ADJG* values are 97 percent of the weighted average of the cohort Ginis.

of lifetime human wealth is only 60 percent of what the standard deviation would be if individuals were frozen into a given spot in the income distribution from one year to the next" (p. 559). This condition of zero mobility is, as we have seen, implicit in *ADJG*, and therefore Parsons' data indicate the degree of upward bias likely in *ADJG*. Finally, Parsons estimated the effect of high mobility: "The actual measure (standard deviation of lifetime earnings) is almost three times larger than it would be if earnings in each year were generated by a random draw" (p. 559).

Lee Lillard's longitudinal study of human wealth reinforces these conclusions. Although his National Bureau of Economic Research sample of white males showed less income dispersion than family units, the relative spread which he found among the Gini coefficients has general significance. Lillard's cohort Ginis averaged .28 while the Gini of lifetime earnings was only .19. He states that "Inequality in earnings at any stage of the life cycle for men over 30, as measured by either the coefficients of variation or the Gini coefficient is 50 percent larger than inequality in human wealth. This conclusion is not affected by changes in the discount rate" (p. 49). Note that my estimate of lifetime income inequality for families in 1972 (using *PG*) was .239 while the average of the cohort Ginis was .334 and the Lorenz-Gini .359 (see my 1975 paper). The last two coefficients are 40 and 50 percent higher than *PG*, or alternatively, my *PG* figure and Lillard's Gini of lifetime inequality are both in the range of 67–72 percent of the conventional cross-section Gini coefficients. However, *ADJG* is .324 or 90 percent of *G*. Thus the longitudinal estimates of lifetime inequality offer striking confirmation that *PG* does not understate inequality but yields estimates in the correct range.

These findings also enable us to resolve a question which has long puzzled researchers: why, if life cycle effects are important, do cohort Ginis average 90 percent of the overall Gini? The reason is now clear: the cohort Gini measure reveals what lifetime inequality would be if persons within a cohort were fixed in rank order throughout the life cycle of the cohort; hence it always overstates inequality in societies where significant intracohort mobility exists. Wertz's *ADJG* coefficient, as noted above, shares the same weakness.

There remains one question. Why and how does *PG* yield closer estimates of lifetime inequality than *ADJG* although neither formula explicitly uses mobility data? The answer briefly is this: the investment in human capital and the stochastic process which generate a society's curved age-income profile also generate the variety of individual income profiles which determine the rank-order transition matrix. Larger mean income differences between age cohorts go hand in hand with increased mobility within a cohort as it moves across the parabolic life cycle path. The *ADJG* is invariant with respect to changes in the average age-income profile since mean income differences between cohorts are removed from the actual income differences (see Wertz, equation (2)). But *PG* responds in an appropriate way. Consider a square matrix, with age cohorts listed top and side, showing in each cell the average of the differences resulting from the income pairings. The main diagonal of the matrix will show the within-cohort pairings while the off-diagonal elements represent income pairings across cohorts (see my 1977 paper, p. 522, Table 1c, for a *PG* matrix in expected gain terms). Given a flat age-income profile, the *PG* and *ADJG* matrices will be the same. With a parabolic profile *PG* and *ADJG* will only be alike in the main diagonal; the off-diagonal elements of *PG* will be smaller than *ADJG* for the reason given by Wertz, namely $|A| - |B| \leq |A - B|$. The greater the mean income spread between cohorts, the greater the difference in the terms of the above inequality. The *ADJG* based on the right-hand term is unresponsive to changes in the shape of the life cycle curve and also to mobility; *PG*, however, varies directly with intracohort inequality and inversely with the mean difference between cohorts, and hence inversely with the degree of mobility. This is a valuable attribute in a coefficient with minimal data requirements, and the evidence from longitudinal income studies supports the conclusion that *PG* provides better estimates of lifetime inequality than other Gini formu-

las based on cross-section data. If equity judgments are conditioned by long-term rather than by transitory inequality, then *PG* can also be used as one element of an equity index.

REFERENCES

L. A. Lillard, "Inequality: Earnings vs. Human Wealth," *Amer. Econ. Rev.*, Mar. 1977, *67*, 42–53.

M. Paglin, "The Measurement and Trend of Inequality: A Basic Revision," *Amer. Econ. Rev.*, Sept. 1975, *65*, 598–609.

———. "The Measurement and Trend of Inequality: Reply," *Amer. Econ. Rev.*, June 1977, *67*, 520–31.

D. O. Parsons, "The Autocorrelation of Earnings, Human Wealth Inequality, and Income Contingent Loans," *Quart. J. Econ.*, Nov. 1978, *92*, 661–69.

B. R. Schiller, "Relative Earnings Mobility in the United States," *Amer. Econ. Rev.*, Dec. 1977, *67*, 926–41.

K. L. Wertz, "The Measurement of Inequality: Comment," *Amer. Econ. Rev.*, Sept. 1979, *69*, 670–72.

Monopoly Output under Alternative Spatial Pricing Techniques: Comment

By SONG-KEN HSU*

In an interesting paper in this *Review*, Melvin Greenhut and Hiroshi Ohta (hereafter G-O) claimed that "monopoly outputs are always greater under spatial price discrimination than they are under simple f.o.b. mill pricing" (p. 705). Later on, William Holahan obtained the basic G-O result in a slightly different model. Both postulated linear demand. However, G-O proposed that their result "does not depend on the linear demand assumption" (p. 712). This may mislead some readers, as there are also non-linear demand conditions for which their results would not hold, as stressed below.

I. A Counter Example

The spatial situation, studied by G-O and Holahan, is that there is a one-dimensional spatial market over which a set of identical consumers are continuously and uniformly distributed and in which a spatial monopolist sells a homogeneous product subject to a strictly positive and constant freight. The monopolist's product cost is

$$T(Q) = cQ + F \tag{1}$$

where Q = the total quantity produced (or, the monopoly output)
c = the constant marginal cost
F = the total fixed cost

Now let the demand function, instead of linear, be

$$q(x) = f[p(x)] = \exp[-\beta p(x)] \tag{2}$$

where x = the distance from the monopolist's mill
$p(x)$ = the delivered price pattern over all x
$= m(x) + tx$
$m(x)$ = the f.o.b. mill price pattern quoted by the monopolist over all x
t = the constant freight
β = a positive parameter

For simplicity, suppose that the demand density for the product at all sites is unity. Thus the local revenue at any x received by the seller is nothing but $m(x) \cdot q(x) = m(x) \cdot \exp[-\beta p(x)]$. Moreover, the monopolist's market boundary is b such that $q(b) = 0$. If we impose no exogenous boundary, the market size is finite if and only if $q(b) = 0$ for some $b < \infty$. For our case of a negative exponential demand, $b = \infty$. Alternatively, one may assume some finite market size b_0, externally imposed as Arthur Smithies did.

The total spatial revenue under spatial price discrimination can be written as

$$R[m(x)] = 2\int_0^b \{m(x) \cdot \exp[-\beta p(x)]\}\, dx \tag{3}$$

and the total profit is

$$\Pi[m(x)] = 2\int_0^b \{[m(x) - c] \cdot \exp[-\beta p(x)]\}\, dx - F \tag{4}$$

where b may be either finite or infinite, depending upon our assumption. A function $m(x)$ is chosen to maximize $\Pi[m(x)]$ in equation (4). Since the first derivative of $m(x)$ with respect to x, i.e., $dm(x)/dx$, does not enter the integrand in equation (4), the variation problem is equivalent to an unconstrained classical programming problem. Thus, the first-order condition of the optimization can be written as

$$m^*(x) = c + 1/\beta \tag{5}$$

where $m^*(x)$ denotes the optimal mill price pattern under spatial discrimination. Equation (5) holds no matter whether one imposes an exogenous $b < \infty$ or not. So for an unregu-

*Assistant research fellow, Institute of Economics, Academia Sinica, Taiwan. I am much indebted to Henry Y. Wan, Jr. for valuable suggestions.

lated spatial monopolist he will quote a mill price pattern which is independent of x. One may call it the optimal uniform mill price pattern, a special case under spatial price discrimination.

As for the monopolist under simple f.o.b. mill pricing, equation (4) turns to be

$$(6)\quad \Pi(\overline{m}) = 2\int_0^b \{(\overline{m} - c) \cdot exp[-\beta(\overline{m} + tx)]\}\, dx - F$$

Now a fixed mill price $\overline{m}$, instead of a mill price pattern over all x, is chosen to maximize $\Pi(\overline{m})$ in equation (6). Let $\overline{m}^*$ denote the optimal fixed mill price. The first-order condition of the optimization becomes

$$(7)\quad \overline{m}^* = c + 1/\beta$$

which is exactly the same as the optimal mill price pattern $m^*(x)$ under spatial price discrimination (see equation (5)).

For the unregulated monopolist, its output, i.e., the "free-spatial demand" (see Greenhut, Hwang, and Ohta) can be written as

$$(8)\quad Q[m(x)] = 2\int_0^b f[m(x) + tx]\, dx$$

and, for the simple f.o.b. mill pricing monopolist, it is

$$(9)\quad Q(\overline{m}^*) = 2\int_0^b f(\overline{m} + tx)\, dx$$

Now since $\overline{m}^* = m^*(x)$ for all x, $Q[m^*(x)] = Q(\overline{m}^*)$. Furthermore, $\Pi[m^*(x)] = \Pi(\overline{m}^*)$ and their welfare effects must also be the same. This proves our assertion.

We note that in the literature of spatial economics the negative exponential form of demand has been mentioned by Smithies as well as Benjamin H. Stevens and C.P. Rydell in a somewhat different context.

The upshot: Whether outputs of spatial monopoly differ under alternative pricing policies is a matter dependent upon the specific form of the demand function.

REFERENCES

M. L. Greenhut and H. Ohta, "Monopoly Output under Alternative Spatial Pricing Techniques," *Amer. Econ. Rev.*, Sept. 1972, *62*, 705–13.

———, **M. Hwang, and H. Ohta,** "Observations on the Shape and Relevance of the Spatial Demand Function," *Econometrica*, July 1975, *43*, 669–82.

W. L. Holahan, "The Welfare Effects of Spatial Price Discrimination," *Amer. Econ. Rev.*, June 1975, *65*, 498–505.

A. Smithies, "Monopolistic Price Policy in a Spatial Market," *Econometrica*, Jan. 1941, *9*, 63–73.

B. H. Stevens and C. P. Rydell, "Spatial Demand Theory and Monopoly Price Policy," *Papers Reg. Sci. Assn.*, 1966, Vol. 17, 195–204.

Monopoly Output under Alternative Spatial Pricing Techniques: Reply

By M. L. Greenhut and H. Ohta*

Our 1972 paper originally proposed that discriminatory outputs are necessarily greater than nondiscriminatory outputs under conditions of spatial monopoly. This proposition was proved under a linear demand function and in a sense generalized in the same paper by intuitive speculation which indicated that the result would hold for non-linear demand cases as well. That speculation was confirmed later in our 1975 book, where we proved that it *holds for all non-linear demands of the particular form given by*

$$p = a - bq^{\alpha} \quad a, b, \alpha > 0 \tag{1}$$

The proof we entered on the basis of (1) does not, however, cover all *conceivable* demand forms. For this reason, our original proposition remains limited.

Song-ken Hsu contends more selectively that our 1972 argument was *misleading.* While we fully agree that any proposition can mislead, we do not accept his counterexample to our own proposition. The fact is that he inadvertently selected an example which does *not* disprove our claim. He assumed the demand form

$$f(p) = \exp[-\beta p], \beta > 0 \tag{2}$$

But a consumer preference pattern of the order of (2) would not provide any incentive at all for a firm to price discriminatorily over its spatial market area. Equation (2) demand is therefore irrelevant and inapplicable to our claim that discriminating pricing generates greater outputs than does f.o.b. pricing. To appreciate the irrelevance of (2), consider the marginal revenue MR, which is derivable from it, namely:

$$(2') \qquad p = f^{-1}(q) = -\frac{1}{\beta} \ln q$$

$$pq = -\frac{q}{\beta} \ln q$$

$$MR = p - \frac{1}{\beta}$$

The monopolist with no constraints on his spatial price policy equates each spatial submarket's marginal revenue, as given by (2′), with his marginal production (c) and transportation cost (t). This process establishes

$$p - \frac{1}{\beta} = c + t \qquad t \geq 0 \tag{3}$$

The profit-maximizing mill price (m) in any submarket is accordingly:

$$m(t) \equiv p - t = c + \frac{1}{\beta} \tag{4}$$

which is the result that Hsu derived.[1]

But what indeed has been demonstrated? Nothing! Hsu has simply provided a demand form which does not generate price discrimination. His demand form yields only a unique f.o.b. price. And so his claim that output and other welfare effects of discriminatory pricing are identical to those of nondiscriminatory pricing is spurious. Phrased otherwise, a valid counterexample to our proposition must be based on a well-defined demand which

*Professor, Texas A&M University, and associate professor, Aoyama Gakuin University and University of Houston (adjunct), respectively.

[1]The c in (4) is not required to be a constant. Instead, it may be assumed to be a function of the firm's total output $Q = \int_0^{\infty} f(m + x)\, dx$, which in turn is a function of mill price m. It follows accordingly that the equilibrium level of c in (4) is simultaneously determined along with Q and m.

provides a firm with incentive to discriminate spatially in the price it charges. For reasons disclosed in our 1975 book (chs. 4, 5), *and not altered by Hsu*, we thus will continue to maintain our original proposition without modification, that is until a meaningful counterexample can be constructed. Hsu's present argument is interesting. But it only proposes a one-sided price practice, and accordingly neither limits our thesis nor in any way supports his extreme charge that it misleads.

REFERENCES

Melvin L. Greenhut and Hiroshi Ohta, "Monopoly Output under Alternative Spatial Pricing Techniques," *Amer. Econ. Rev.*, Sept. 1972, *62*, 705–13.

——— **and** ———, *Theory of Spatial Pricing and Market Areas*, Durham 1975.

S.-k. Hsu, "Monopoly Output under Alternative Spatial Pricing Techniques: Comment," *Amer. Econ. Rev.*, Sept. 1979, *69*, 678–79.

Constant-Utility Index Numbers of Real Wages: Comment

By P. J. LLOYD*

Paul Samuelson and Subramanian Swamy in their survey of index-number theory in this *Review* emphasized that "The fundamental point about an economic quantity index, which is too little stressed by writers, Leontief and Afriat being exceptions, is that it must itself be a cardinal indicator of ordinal utility" (p. 568). In a later article in this *Review* John Pencavel has endeavored to compute real wage indices in this sense. He interprets each of his indices as an "index of the individual's welfare" (p. 93). His two series of real wages are derived from an estimated indirect Stone-Geary utility function which incorporates nonlabor income of the wage earners and an endogenous work-leisure choice. In one series the increase in real wages over the period 1934–67 was substantially less than the index of money wages deflated by the Consumer Price Index or the Bureau of Labor Statistics series of real spendable weekly earnings of production workers, whereas in the second series the increase was substantially greater than in these other series over the same period. He has also constructed an index of real nonlabor income. My contention is that none of these indices is a true quantity index, but a genuine true quantity index can be obtained from the indirect utility function by using a slightly different definition of income. Moreover, this can be done for any regular utility function. For a family of functions which includes the Stone-Geary, this index is equal to an index of deflated incomes and is the canonical dual of the true price index.

For any utility function one can obtain a quantity index of real income across income-price situations directly and simply by taking the ratio of the indirect utility function in period t to that at the situation in a base period 0. That is,

$$(1) \quad Q_t = V(p_{t1}, \ldots, p_{tn}, w_t, y_t) / V(p_{01}, \ldots, p_{0n}, w_0, y_0)$$

The symbols are the same as in Pencavel. When the direct utility function is taken to be Stone-Geary,

$$(2) \quad U(x_1, \ldots, x_n, l) = \sum_{i=1}^{n} B_i \ln (x_i - \gamma_i) + \theta \ln (l - \gamma_l)$$

the dual indirect utility function is

$$(3) \quad V(p_1, \ldots p_n, w, y) = \prod_{i=1}^{n} (B_i/p_i)^{B_i} \cdot (\theta/w)^{\theta} (y + w\gamma_h - \sum_{i=1}^{n} p_i \gamma_i)$$

Substituting in equation (1) and rearranging, one obtains the true quantity index of real incomes (or, more loosely, real wages)

$$(4) \quad Q_t = \frac{Y_t / \prod_i p_{ti}^{B_i} w_t^{\theta}}{Y_0 / \prod_i p_{0i}^{B_i} w_0^{\theta}}$$

$$Y_t = (y_t + w_t T) - (\sum_{i=1}^{n} p_{ti}\gamma_i + w_t \gamma_l)$$

In this case one finds that the index can be obtained by taking in each period *total* income (which is equal to nonlabor income plus the potential labor income) less expenditure on the minimum bundle, and deflating by the appropriate geometric measure of the price level. This price level includes the price of the leisure commodity. The weights B_i and θ are the shares of the budget excluding the minimum purchases of all commodities. In fact, this price level is the minimum cost in period t of the above-subsistence bundle required to attain the level of utility of period 0.

*Senior fellow, economics department, Research School of Pacific Studies, Australian National University.

However, Pencavel did not proceed in this way. He defines a "constant-utility" wage rate for year t as that wage rate w_t^* such that, with the prices and nonlabor income as given in this year, the utility of the wage earner would be equal to that of a chosen base year. That is, using the indirect utility function, w_t^* is defined implicitly by

$$(5)\quad \left(y_t + w_t^*\gamma_h - \sum_i p_{ti}\gamma_i\right)/\prod_i p_{ti}^{B_i} w_t^{*\theta} = \left(y_0 + w_0\gamma_h - \sum_i p_{0i}\gamma_i\right)/\prod_i p_{0i}^{B_i} w_0^{\theta}$$

The series of the ratio (w_t/w_t^*) is then taken as one index of a true real wage rate. Since this index incorporates the effects of changes in nonlabor income between the base period (y_0) and period t (y_t), Pencavel presents another preferred index in which the nonlabor income is held constant at the base period level. This w_t^{**} is defined implicitly by

$$(6)\quad \left(y_0 + w_t^{**}\gamma_h - \sum_i p_{ti}\gamma_i\right)/\prod_i p_{ti}^{B_i} w_t^{**\theta} = \left(y_0 + w_0\gamma_h - \sum_i p_{0i}\gamma_i\right)/\prod_i p_{0i}^{B_i} w_0^{\theta}$$

It gives the second index (w_t/w_t^{**}).

Neither (w_t/w_t^*) nor (w_t/w_t^{**}) is a true quantity index of real income. The former does have the property that $(w_t/w_t^*) \gtreqless 1$ as $V_t/V_0 \gtreqless 1$ but the latter index does not. This is evident from the definition. Holding constant y and obtaining w_t^{**}, the sign of the change of actual utility between periods 0 and t as given by $V(p_t, w_t, y_t)$ depends on the actual value of nonlabor income in period t. The index (w_t/w_t^*) does not necessarily rank three or more situations in the same order as V, even if there is no nonlabor income in these periods.[1] It cannot, therefore, be a true quantity index.

It is possible to obtain a true quantity index in the indirect manner of Pencavel by first obtaining the ratio of two expenditures required to maintain utility and then dividing the actual income by the numerator of this ratio. By analogy with Pencavel's other measures of wage income, I define

$$Y_t^* = cU_0 \prod_i p_{ti}^{B_i} w_t^{\theta}$$

where

$$c = \prod_i B_i^{B_i}\theta^{\theta}$$

This is the minimum cost, at prices in situation t, of purchasing the level of utility, U_0. Further, on the assumption of utility maximization,

$$cU_0 \prod_i p_{0i}^{B_i} w_0^{\theta} = Y_0$$

Rearranging equation (4)

$$(7)\quad Q_t = \frac{Y_t/Y_0}{Y_t^*/Y_0}$$

$$(8)\quad = Y_t/Y_t^*$$

Unfortunately, Pencavel did not include the correct choice of income expenditure concepts among the three alternatives which he tried.[2] One must not separate labor and nonlabor incomes since both jointly determine the level of utility. Moreover, it is confusing to call the true quantity index so obtained a "constant-utility" index. Although such a measure may be expressed as the ratio of the index of incomes divided by a constant-utility or true index of prices, a true quantity index must be a measure in commodity space which indexes the change in utility.

Although the choice of a Stone-Geary utility function is arbitrary and limited, it provides a quantity index of real incomes which has interesting properties. It is apparent from equation (4) that the quantity and price indices for this function have the Fisher "factor-reversal" property that their product is equal to the index of "income" when this is

[1]The actual relationship between (w_t/w_t^*) and V_t/V_0 is quite complex and depends on all commodity prices and nonlabor incomes as the reader may see by comparing equation (4) with (w_t/w_t^*) where w^* is defined as in equation (5).

[2]In fn. 4 Pencavel comes close to the true quantity index when he contemplates an index of "full income" which is total income including potential labor income, that is, $(y_t + wT)$. If one adjusts this further by subtracting the expenditure on the minimum bundle one obtains the true index given in the text. Apart from overlooking the expenditure on minimum bundle, Pencavel rejected his full income measure because the maximum work time T is arbitrary. However, the concept of a fixed minimum work time γ_l, which is also required in the calculation of utility maximization and expenditure minimization is equally arbitrary.

defined appropriately. Moreover, the price index is independent of the reference level of utility and the quantity index is independent of the reference level of prices.

In fact, when one recognizes that the Stone-Geary function is homothetic to the point $\gamma = (\gamma_1, \ldots, \gamma_n, \gamma_l)$, one can extend the Samuelson-Swamy duality theorems and obtain some results which are useful for the calculation and interpretation of true indices of real income. For this purpose I define a utility function $U(x)$ as a member of the family of generalized homothetic functions if

$$(9) \quad U(x) = T(g(x-\gamma)) \qquad x = (x_1, \ldots, x_{n+1})$$

where g is homogeneous of degree $+1$, T is a continuous nonnegative and nondecreasing transformation and γ is a vector of constants.[3] It is convenient to redefine leisure as the $(n+1)$ commodity and use vector notation. In addition to the Stone-Geary, this family includes such utility functions as the two-stage *CES* function used by Murray Brown and Dale Heien and the homogeneous indirect translog function used by Marilyn Manser. With this definition and the corresponding definition of minimum expenditure as the minimum expenditure on all above-subsistence quantities, and using the canonical linearly homogeneous function with $T = 1$ to represent consumer preferences, the duality theorems which Samuelson and Swamy proved for homothetic functions carry over to generalized homothetic functions.

This dual quantity index can be interpreted in three ways.

PROPOSITION: *If and only if the utility function is a generalized homothetic function the quantity index can be stated in three equivalent ways:*

(i) *The Samuelson-Swamy Index*: the ratio of minimum expenditures, excluding expenditures on the minimum or subsistence quantities, required to attain the levels of utility in the two situations, viz.,

$$(10) \qquad Q_t = e(x_t)/e(x_0)$$

(ii) *The Deflated Income Index*: the ratio of the levels of deflated income, viz.,

$$(11) \qquad Q_t = [Y_t/\epsilon(p_t)]/[Y_0/\epsilon(p_0)]$$

(iii) *The Malmquist Index*: the quantity λ such that the

$$(12) \qquad g(\lambda(x_0-\gamma)) = g(x_t-\gamma) = U_t$$

In (i) e is the minimum expenditure function redefined to exclude expenditure on the minimum bundle. In (ii) $\epsilon(p)$ is the dual linearly homogeneous price function. In (iii) the Malmquist index can be shown to be equal to the ratio of two values of the minimum distance function (see Ronald Shephard, ch. 3), measuring distance from the point γ.

PROOF:

Consider (ii): Q_t is defined as the ratio of two values of the indirect utility function in its canonical form (V_t/V_0). This dual direct utility function is linearly homogeneous in the variables $(x-\gamma)$. Shephard has proven that the indirect utility function has the form

$$(13) \qquad V = y'/\epsilon(p)$$

where y' is total income and $\epsilon(p)$ is a linearly homogeneous price function if and only if $U(x)$ is linearly homogeneous. This proposition carries over to the canonical form of the generalized homothetic function by simply redefining variables as I have done.

Consider the definition of the index suggested by Sten Malmquist in (iii). If and only if $U(x)$ is linearly homogeneous

$$g(\lambda(x_0-\gamma)) = \lambda g(x_0-\gamma) = U_t$$

Hence, $\lambda = g_t/g_0$ which is equivalent to (ii).

Consider the Samuelson-Swamy definition of (i). Again measuring expenditure from the minimum consumption point, it follows

[3]It is also assumed that g is strictly quasi concave and has strictly positive first derivatives. These restrictions ensure that the commodity demand functions and the indirect utility function are single valued and that the utility-maximizing bundle lies on the boundary of the budget set. They rule out functions such as the quadratic utility function which satisfy equation (9) but have linear Engel curves converging to a maximum point. If strict quasi concavity and nonsatiation are satisfied the results can also be applied locally to such functions.

immediately from the linear homogeneity of $g(x - \gamma)$ and the consequential linearity of the expansion paths through this point that the ratio of these expenditures is equal to the ratios of consumption and utility as in (iii) and (ii).

The equivalence of (i) and (ii) provides the necessary and sufficient condition for deriving an index of real income by the natural process of deflating an index of money income by a true price index which is independent of the reference level of utility, namely, generalized homothetic.

If the utility function is not generalized homothetic and $U(x)$ is defined over the whole nonnegative orthant of commodity space, the Malmquist Index is greater than the Samuelson-Swamy Index and the latter index is not independent of the reference price situation; that is, it is not independent of the point on U_0 from which one starts. One may still obtain a Samuelson-Swamy true quantity index by dividing total income by a true price index if one combines either p_0 as the reference price level with U_t as the reference utility level for the quantity and price index, respectively, or one combines p_t with U_0.

REFERENCES

M. Brown and D. Heien, "The S-Branch Utility Tree: A Generalization of the Linear Expenditure System," *Econometrica*, July 1972, *40*, 737–47.

S. Malmquist, "Index Numbers and Indifference Surfaces," *Trabajos de Estadistica*, No. 4, 1953, *4*, 1–42.

M. E. Manser, "Elasticities of Demand for Food: An Analysis Using Non-additive Utility Functions Allowing for Habit Formation," *Southern Econ. J.*, July 1976, *43*, 879–91.

J. H. Pencavel, "Constant-Utility Index Numbers of Real Wages," *Amer. Econ. Rev.*, Mar. 1977, *67*, 91–100.

P. A. Samuelson and S. Swamy, "Invariant Economic Index Numbers and Canonical Duality: Survey and Synthesis," *Amer. Econ. Rev.*, Sept. 1974, *64*, 566–93.

Ronald W. Shephard, *Theory of Cost and Production Functions*, Princeton 1970.

Constant-Utility Index Numbers of Real Wages: Reply

By JOHN H. PENCAVEL*

I appreciate Peter Lloyd's thoughtful comment on my "Real Wages" paper and I offer some observations below on his proposed Q_t index. First, however, I should point out that Lloyd's remarks are not directly relevant to the main issues with which that paper was concerned. The paper was motivated as an examination of the published Bureau of Labor Statistics' series on movements in real wage rates and in real earnings. It was similar in spirit to the studies evaluating the fixed-weight Consumer Price Index by comparing its movements with those of a constant-utility index number of the cost of living.[1] The index numbers reproduced in Table 1 of my paper and which are employed in the regression analysis later in the paper make no claim to include income from sources other than the sale of a worker's labor time. And, indeed, such a separation of labor from nonlabor income is required for an analysis of a wage rate index. By determining the constant-utility wage rate, I was making use of a *price* (or wage) solution to the indirect utility function as distinct from the income solution. Analogous to the income solution, the wage (or price) solution also may be identified as the outcome of an extremum problem. In particular, it represents the maximum wage rate permissible, given nonlabor income and all other prices, such that the maximum utility attainable is not greater than a specified level.[2] In short, though some of the language in the paper may have suggested otherwise, I did not construct nor did I intend to construct a "true" quantity index of real income from all sources.

If I had wanted a true quantity index of this sort, at the time of writing the paper, I would have drawn a parallel with the conventional (zero endowments) model of consumer demand by using that expenditure function which corresponds to the "full income" solution of the indirect utility function. In this context, full income means $y + wT$ where y represents nonwage income, w the (assumed exogenous) wage rate, and T the total amount of time to be allocated between market and nonmarket activities. For a utility function of the Stone-Geary type, the full income required in period t to attain the level of utility enjoyed in period 0 is given by

$$F_t^* = \sum_{i=1}^{n} p_{ti}\gamma_i + w_t\gamma_l + \prod_{i=1}^{n}\left(\frac{p_{ti}}{p_{0i}}\right)^{B_i}\left(\frac{w_t}{w_0}\right)^{\theta}\left(y_0 + w_0\gamma_h - \sum_{i=1}^{n} p_{0i}\gamma_i\right)$$

where the notation is the same as in the original paper. A natural index of the consumer-worker's overall welfare is provided by the ratio of observed full income (F_t) to F_t^*. I explicitly entertained this notion, but I was reluctant to implement it because of the difficulty of giving quantitative expression to T.[3] It is necessary in this instance since, as is evident from the previous equation, F_t^* involves γ_l and γ_h separately (where $T = \gamma_l + \gamma_h$). I could have not worried about this problem and followed Gary Becker by assigning a value of 8,736 hours per year to T; or, if I had adopted H. Gregg Lewis' assumption, T would have been set at 100 hours per week (or approximately 5,200 hours per year); or I could have fixed T at 576 hours per month (or 6,912 hours per year) as does Roger Betancourt. The use of each of these values of T would have produced different (and, I conjec-

*Associate professor of economics, Stanford University.

[1] A summary of most of these studies appears in Jack Triplett.

[2] The properties of a price or wage rate solution of the indirect utility function are contained in the paper by Lawrence Lau and myself.

[3] In my "Revised Estimates," I did construct something similar to this by interpreting the estimated γ_h as the maximum feasible hours of work and by forming $(y_t + w_t\gamma_h)^*$.

ture, substantially different) estimates for a full income welfare index.

Lloyd's ingenious proposal is to use his equation (8) as a quantity index of overall welfare, namely,

$$Q_t = \frac{Y_t}{Y_t^*} = \frac{y_t + w_t\gamma_h - \sum_{i=1}^{n} p_{ti}\gamma_i}{\left[\prod_{i=1}^{n} B_i p_{ti}(x_{0i} - \gamma_i)\right]^{B_i} [\theta w_t(\gamma_h - h_0)]^{\theta}}$$

The numerator of this expression measures maximum discretionary income in period t. The denominator is a geometric weighted average of commodity expenditures and labor income involving period 0's quantities valued at period t's prices; or, equivalently, the cost of attaining period 0's welfare at period t's prices. The appealing feature of Lloyd's Q_t that he neglected to mention is that, unlike the index F_t/F_t^*, it does not require any assumption with respect to T. That is, all the parameters in Q_t fall out of the estimation process and its construction requires no prior assumption concerning the amount of time to be divided between market and nonmarket activities. For this reason, I applaud Lloyd's proposed Q_t index.

REFERENCES

G. S. Becker, "The Allocation of Time Over the Life Cycle," in Gilbert R. Ghez and Gary S. Becker, eds., *The Allocation of Time and Goods Over the Life Cycle*, New York 1975.

R. R. Betancourt, "The Estimation of Price Elasticities from Cross Section Data Under Additive Preferences," *Int. Econ. Rev.*, June 1971, *12*, 283–92.

L. J. Lau and J. H. Pencavel, "Properties of the Wage Solution of the Indirect Utility Function," mimeo., Stanford Univ., Apr. 1978.

H. G. Lewis, "Economics of Time and Labor Supply," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 29–34.

P. J. Lloyd, "Constant-Utility Index Numbers of Real Wages: Comment," *Amer. Econ. Rev.*, Sept. 1979, *69*, 682–85.

J. H. Pencavel, "Constant-Utility Index Numbers of Real Wages," *Amer. Econ. Rev.*, Mar. 1977, *67*, 91–100.

———, "Constant-Utility Index Numbers of Real Wages: Revised Estimates," *Amer. Econ. Rev.*, Mar. 1979, *69*, 240–43.

J. E. Triplett, "The Measurement of Inflation: A Survey of Research on the Accuracy of Price Indexes," in Paul H. Earl, ed., *Analysis of Inflation*, Lexington 1975.

Valuing Market Benefits and Costs in Related Output and Input Markets

By DENNIS W. CARLTON*

A frequent application of cost-benefit analysis is to measure the cost of a tax or the benefit of a technological improvement in some input market. In many instances data on the directly affected input market may not be available, but data on output markets or other input markets may be available. For example, to evaluate a tax on fertilizer, one would like data on the amount of fertilizer purchased. Such data may be unavailable, while data on bushels of wheat, tractors, or land purchased may be readily available. An important question is how well these related markets reflect costs and benefits of the directly affected input market. Being able to use data from related markets to measure benefits or costs can enable the analyst to overcome many data problems and thereby widen the applicability of his tools.

This paper argues that related output-market measures can perfectly reflect an input-market cost or benefit while related input-market measures can reflect this cost or benefit under special though plausible conditions. Section I proves that input-market costs and benefits are totally reflected in the output market for the general case when supply curves slope upwards (i.e., factors earn rents). This result was first proved by Daniel Wisecarver for the special case of two inputs, and subsequently by James Anderson and by Richard Schmalensee for the special case of perfectly elastic supply curves. Section I emphasizes that the output-market measure can enable an analyst to perform precise measurements of social costs even when data on the directly affected input market are unavailable. Sections II and III discuss how information from other observable input and output markets can be used to measure costs and benefits that arise in an unobservable market.

I

I wish to show that an input-market distortion can be completely measured in the output market.[1] Without loss of generality, let us examine the case where a tax is placed on only one factor, and assume inputs used to produce the output can be used to produce only this output. An obvious generalization of the proof establishes it for the case where many inputs are taxed and many outputs can be produced by the inputs. I assume competition in all markets.[2] I do not require that factor supply curves be infinitely elastic and instead allow for the case where factor supply curves slope upwards. This means that factors can earn rents. There are two standard reasons why factor supply curves can slope upwards: the opportunity cost of the factor may not be constant as in the labor supply decision; second the marginal cost of producing the factor may be rising as, for example, in the case of mineral deposits of different extraction costs, or in the case of agricultural land that requires different amounts of fertilizer.

Let $r_i(t)$ be the equilibrium wage *received* by factor i when the tax on factor 1 is t. Therefore, a firm pays $r_1(t) + t$ for factor 1, $r_2(t)$ for factor 2, etc. Let $X_i(t)$ be the amount of factor i demanded in equilibrium when the tax on factor 1 is t. The deadweight loss S^I of an increase in the tax on factor 1 from t_0 to t_1 is measured in the input market by

*University of Chicago. I thank Arnold C. Harberger for encouraging me to write up these results and for perceptive comments. I also thank Kent Kimbrough, John Panzar, Russell Roberts, Robert Willig, the referee, and George Borts for helpful comments, and the *NSF* for research support.

[1]It is also true that an input-market benefit is completely reflected in the output market. This result follows from a proof almost identical to the one given below.

[2]The method of proof is similar to that used by Schmalensee (1976).

$$S^I = \int_{X_1(t_1)}^{X_1(t_0)} [w_1^d(X_1) - w_1^s(X_1)]dX_1$$

where $w_1^d(X_1)$ = the demand wage for X_1 units of factor 1
$w_1^s(X_1)$ = the supply wage for X_1 units of factor 1.[3]

Notice that

$$(1)\quad \frac{dS^I}{dt_1} = -[w_1^d(X_1(t_1)) - w_1^s(X_1(t_1))] \cdot \frac{dX_1}{dt} = -t_1\frac{dX_1}{dt}$$

I now show how the measurement of deadweight loss in the output market will lead to exactly the same result as (1), provided we measure costs appropriately. Let the equilibrium output when taxes on factor 1 are t be denoted by $Q(t)$. The loss from raising the tax on factor 1 from t_0 to t_1 can be measured as the change in the area under the output demand curve minus the change in "true" costs (i.e., excluding rent and tax revenue changes) of producing the output.

Let the costs (excluding tax revenue) of producing output $Q(t)$ be denoted by[4]

$$C(t) \equiv \sum_i r_i(t)X_i(t)$$

The change in $C(t)$ as the tax rate increases is

$$C'(t) = \sum r_i(t)\frac{dX_i}{dt}(t) + \sum\frac{dr_i}{dt}X_i(t)$$

The change in cost relevant for calculations of welfare loss is not $C'(t)$ since $C'(t)$ includes the term $\Sigma(dr_i/dt)X_i(t)$ which is the change in *rents* that accrue to factor owners as taxes are altered. The cost change relevant for the cost-benefit calculation excludes rent and equals

$$(2)\quad \overline{C}'(t) \equiv C'(t) - \sum\frac{dr_i}{dt}X_i(t) = \sum r_i(t)\frac{dX_i}{dt}(t)$$

Denote the area under the (inverse) output demand curve, $P(Q)$, between the quantities corresponding to the tax rates t_0 and t by

$$CS(t) \equiv \int_{Q(t)}^{Q(t_0)} P(Q)dQ$$

hence

$$\frac{dCS}{dt} = -Q'(t)\,P(Q(t))$$

The measure of the input-market distortion of raising the tax from t_0 to t_1 when measured in the output market is the sum of the changes in cost and in the area under the output demand curve, and equals

$$S^0 = \int_{t_0}^{t_1}\left[\frac{dCS}{dt} + \overline{C}'(t)\right]dt$$

Since the output and factor markets are competitive, we have that

$$Q'(t) = \sum_i MPP_i\frac{dX_i}{dt} = \left[\sum r_i(t)\frac{dX_i}{dt} + t\frac{dX_1}{dt}\right]\frac{1}{P(Q(t))}$$

where MPP_i is marginal physical of factor i and where we have used the relation that in competitive markets, the wage equals the value of the marginal physical product. Inserting $Q'(t)$ into the expression for dCS/dt, and using the expressions for S^0 and $\overline{C}'$, we obtain

$$\frac{dS^0}{dt_1} = -\sum r_i(t_1)\frac{dX_i(t_1)}{dt} - t_1\frac{dX_1(t_1)}{dt} + \sum r_i(t_1)\frac{dX_i(t_1)}{dt}$$

or

$$\frac{dS^0}{dt_1} = -t_1\frac{dX_1}{dt}$$

which is the same as dS^I/dt_1 in (1). We therefore conclude that S^I and S^0 yield identical measures of the distortion.

It is useful to illustrate diagrammatically how one could calculate deadweight loss in an output market. Suppose that the tax on factor 1 is increased from 0 to some small ϵ. Let us

[3]Whenever cost-benefit analysis is done, compensated demand and supply curves must be used. The input demand curves are the demands for input that arise when the amount of output produced is given by a compensated demand. With a non-linear production-possibility frontier, the relevant curves to use are compensated "general equilibrium" demand curves which incorporate price changes in other markets into the demand curve. See Harberger and Peter Diamond and Daniel McFadden.

[4]For simplicity of notation, the cost of producing output $Q(t)$ is not written as $C(Q(t))$ but as $C(t)$.

first calculate what the relevant cost change ($\Delta\overline{C}$) equals. By definition (see (2)) $\Delta\overline{C}$ will approximately equal $\overline{C}'(0)\epsilon$, or the difference in total costs after taxes and rents have been excluded. Mathematically,

$$\Delta\overline{C} = (TC_\epsilon(Q_\epsilon) - RENT_\epsilon(Q_\epsilon)) - (TC_0(Q_0) - RENT_0(Q_0)),$$

where Q_0 = output when tax is 0
Q_ϵ = output when tax is ϵ
$TC_0(Q)$ = total cost of producing Q when tax is zero
$TC_\epsilon(Q)$ = total cost of producing Q when tax is ϵ (does not include tax payments)
$RENT_0(Q)$ = rent earned when Q is produced and tax is 0
$RENT_\epsilon(Q)$ = rent earned when Q is produced and tax is ϵ

Rewrite the above expression for $\Delta\overline{C}$ as

$$(3)\quad \Delta\overline{C} = [TC_\epsilon(Q_\epsilon) - RENT_\epsilon(Q_\epsilon)] - [TC_0(Q_\epsilon) - RENT_0(Q_\epsilon)] + [(TC_0(Q_\epsilon) - RENT_0(Q_\epsilon)) - (TC_0(Q_0) - RENT_0(Q_0))]$$

Let $C_0(Q)$ be the supply (unit cost) curve when the tax is zero and $C_\epsilon(Q)$ be the supply (unit cost) curve when the tax on factor 1 is ϵ and when tax revenues have been subtracted from costs. $C_\epsilon(Q)$ lies above $C_0(Q)$ because the tax causes output to be produced with inefficient factor proportions. Notice that the first term in (3), $[TC_\epsilon(Q)_\epsilon - RENT_\epsilon(Q_\epsilon)]$, is given exactly by the area beneath the $C_\epsilon(Q)$ curve between 0 and Q_ϵ.[5] Similarly the second term in (3), $[TC_0(Q_\epsilon) - RENT_0(Q_\epsilon)]$, is given by the area beneath the $C_0(Q)$ curve between 0 and Q_ϵ. The difference between these first two terms is then the area between the $C_\epsilon(Q)$ and $C_0(Q)$ curves from 0 to Q_ϵ. The last expression in (3) is nothing more than the incremental social cost of producing $Q_\epsilon - Q_0$, which approximately equals $[Q_\epsilon - Q_0]\,C_0(Q_0)$.[6]

Using the above analysis we can decompose

FIGURE 1

$\Delta\overline{C}$ into the difference between two areas in Figure 1. The first is a banana-shaped area representing the difference between the first two expressions in (3). The banana is the increased cost of production caused by input distortions that arise from the tax on the factor. The second area is a box that represents the last term in (3). The box represents the decrease in cost that comes from a reduction in output. By construction, the difference between the banana and the box is precisely what $\overline{C}'(0)\Delta t$ measures for $\Delta t = \epsilon$ in (2).

If we let $(\epsilon_1, \epsilon_2, \ldots, \epsilon_\infty)$ be a sequence whose sum converges to t, and if we let $C_t(Q)$ be the supply (unit cost) curve when the tax on factor 1 is t and taxes are excluded from costs, then we can build up the area relevant for consideration of the cost side for a tax t on factor 1 that reduces output from Q_0 to Q_t. That area is given as the sum of all the infinitesimal bananas minus boxes as $\Sigma_i\epsilon_i \rightarrow t$. It equals $ABCD$ minus $DEFG$ in Figure 2. In Figure 2 the curve $C_t^*(Q)$ is the unit cost curve including tax revenues when the tax on input 1 is t. The demand curve is denoted as $Q(p)$. The intersection of $C_t^*(Q)$ and $Q(p)$ at H determines the post tax quantity of output Q_t. The area under the demand curve "lost" as a result of the tax is $HEFG$. The net loss equals $ABCD + HEFG - DEFG = ABCD + HGD$ and is illustrated as the shaded area in Figure 2. The tax revenues are given by the area of the rectangle $HIJC$.

[5]This result makes use of the well-known fact that the rents earned can be calculated as the area between the supply curve and the horizontal price line.

[6]Notice that the supply (unit cost) curve reveals the marginal cost as well as the average cost at each output level.

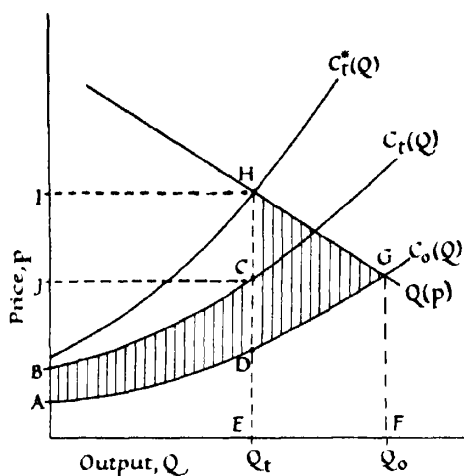
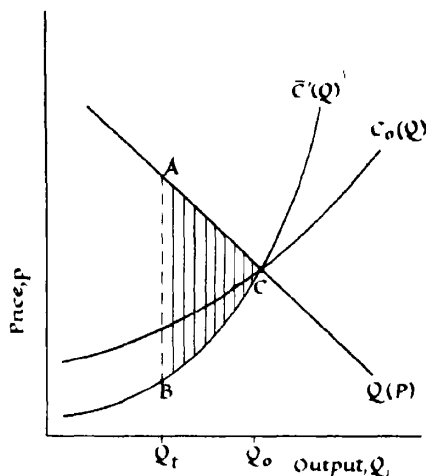



FIGURE 2

FIGURE 3

Since $\bar{C}'$ was defined as the "relevant" area (i.e., a banana minus a box), (see (2) and (3)), an alternative diagram to use in the benefit-cost analysis is Figure 3 above.[7] The dead-weight loss is given as a "triangle" ABC. The reason this diagram is not as useful as Figure 2 is because $\bar{C}'$ is a complicated function (see (2)) that must be derived from the information contained in the cost curves of Figure 2.

The rather complicated construction of loss in Figure 2 should be contrasted with the simple determination of loss in the directly affected factor market. From our previous proof we are guaranteed that the shaded area in Figure 2 equals the familiar triangle ABC of Figure 4, where S and D in Figure 4 refer to the supply and demand for the factor on which the tax is placed. The tax revenues in Figure 4 are given by the area of the rectangle $ABDF$.

When information about supply of all inputs, prices of all inputs, and demand for output are available, then there is no reason to prefer using Figure 4 or Figure 2. On practical grounds, Figure 4 may often be the preferred alternative since its calculation requires knowledge only of one supply and demand curve. Information on other factor markets or on the output market is not needed.

Is it ever the case that information on the affected input market is so poor that Figure 2 can be used but not Figure 4? The answer is yes, and this result provides the sole justification for trying to measure input distortions (or benefits) in output markets.[8] A simple example is the easiest way to illustrate this point.

Suppose that an undistorted output market is observed for a long time. Observable shifts in exogenous (for example, income, weather) variables enable the analyst to identify the relevant pretax supply and demand curves.

[7]Here we draw $\bar{C}'$ as a function of output. The notation in Figure 1 is the same as in Figure 2.

[8]This statement appears to conflict with Wisecarver's belief that "practical measurement of the substitution compenent in the output market is intractable" (p. 367) and that output measures "do not fit naturally into the framework of applied welfare economics" (p. 367). Wisecarver is, however, correct to insist on the simpler input market measure *when* it is available. The recent paper by John Panzar and Robert Willig also *appears* to imply that only direct input-market measures are correct. However, a careful reading of their paper shows that this implication would be a misrepresentation of their views. Panzar and Willig argue correctly that consumer's surplus in input and output markets will differ, but do not discuss how taking account of a banana-shaped area (that represents increased production costs caused by factor-market distortions) can enable use of output-market measures. This paper is in agreement then with Panzar and Willig. Without taking the banana into account, it is fruitless to perform cost-benefit analysis in the output market.

FIGURE 4

Data on certain input prices and quantities demanded are not available.[9] A tax is imposed on the input[10] and the policy remains in effect for several years. Observable shifts in exogenous variables enable the analyst to identify the posttax supply curve. Information on taxes collected $T(Q)$ is available. If information on pretax factor prices and quantities is unavailable, then the distortion cannot be measured in the input market. However, using Figure 2, the distortion can readily be measured in the output market. We let $C_0(Q)$ be the pretax supply curve, $C_r^*(Q)$ be the posttax supply curve, and $C_r(Q)$ be $C_r^*(Q) - (T(Q)/Q)$. The shaded area in Figure 2 gives the correct measure of deadweight loss.

Measuring an input-market distortion in an output market is more tedious than measuring the distortion in the directly affected input market. However, it may be possible to use the output-market distortion measure in cases where data unavailability precludes the use of the input-market distortion measure. In any practical application one must keep in mind the usual caveats about using compensated demand curves and "reduced-form general equilibrium" demand curves (see Harberger and Diamond-McFadden) and giving proper attention to distributional issues. Moreover, if inputs are used to produce *several* outputs then a tax on an input(s), when not measured in its own market(s), must be measured in *all* the affected output markets. The need to look at several markets when output distortion measures are used emphasizes the simplicity of using input distortion measures when data permit. However, when data are lacking, the more complicated output-market distortion measures can be used to accurately reflect costs and benefits in input markets. It is this property that should make the technique of measuring input-market distortions or benefits in the output market a powerful and useful one for cost-benefit analysts.

II

Can the analyst ever use information on *other input* markets to evaluate an input-market benefit or distortion? A closely related question is whether the analyst can ever use information on *other output markets* to evaluate an output market benefit or distortion. The answer to both questions is yes but only with special though plausible assumptions. The intuition is that if the price of one input (output) is affected, the demand curves (see fn. 3) for other inputs (outputs) should also be affected. This section analyzes the use of data from other input markets, while Section III analyzes the use of data from other output markets.

If one has prior knowledge about production, then it is sometimes possible that information from only *some* input markets can provide complete information about *all* input markets. The issue is whether the prior plus observable information allows the derivation of the missing data. For example, suppose it is known that output (Q) is produced by a

[9]Input prices may of course change with the quantity of input demanded, but the (unobservable) supply *curve* of inputs must remain stable over time (or else the change must be parameterized in terms of observable variables).

[10]All inputs are assumed to be used to produce only this one output. Otherwise, changes in factor-price ratios will reverberate into other output markets. In such a case, each output market would have to be separately analyzed as in Figure 2, and the total deadweight loss would equal the sum of the deadweight losses across all affected output markets.

Cobb-Douglas technology using two factors, labor L and capital K, according to $Q = F(L, K) = AL^{\alpha}K^{1-\alpha}$. Suppose that factor prices and L are observed and that α is known. Then, knowledge of L together with the ratio of factor prices is sufficient to construct the demand for K. This demand relation can then be used to measure any benefits or distortions in the capital market.

Another class of cases in which cost-benefit analysis can be done on unobservable markets are those in which prior information falls short of identifying the whole production process, but still enables calculation of the desired magnitudes. Suppose that for simplicity we assume constant returns to scale, perfectly elastic factor supplies, and a technological improvement that lowers the cost of input 1 from r_0 to r_1. How can we measure this benefit? First, we could measure it directly in the input market as a trapezoidal area between r_0 and r_1 and the derived demand curve for input 1 (see, for example, the second diagram in Figure 6). Alternatively, from Section I we know that we could measure it in the output market by the area composed of a banana[11] (the difference between the new and old marginal cost curve) plus a triangle representing the additional consumer's surplus. Can we ever use information on other observable *inputs* to derive the relevant measure? The answer is yes—but only under a special though perfectly reasonable assumption. The special assumption is that as the price of the observable input becomes arbitrarily large, the price of the final product rises enough to reduce the value of consumption to zero.[12] Under this assumption it is possible to show that the appropriate difference in areas bounded by the new and old input demand curve for *any* related input can be used to exactly measure the gain to society.

To prove this point suppose that input 1 and input 2 produce output 1. Let the wages of inputs 1 and 2 be denoted by r and s, respectively. Let $C(r, s)$ be the unit cost function for producing output. Let $Q(u, p_1, p_2)$ be the individual's compensated demand curve for output 1 at utility level u facing prices p_1 for output 1 and p_2 for other outputs. Let r fall from r_0 to r_1 because of a technological change. Let s_0 be the wage received by input 2. The direct measure of benefit is given in the input market by[13]

$$(4) \quad -\int_{r_0}^{r_1} Q(u, C(r, s_0), p_2)\frac{\partial C}{\partial r}(r, s_0)\,dr$$

Using the reasoning from the previous section, the indirect measure of benefit measured in the output market is

$$(5) \qquad -\int_{C(r_0,s_0)}^{C(r_1,s_0)} Q(u, p_1, p_2)\,dp_1$$

which immediately reduces to (4) once the change of the variable r for p_1 is performed using the relation $p_1 = C(r, s_0)$.

Now let us look at the demand for input 2. Consider Figure 5 below. The initial derived demand for input 2 shifts in response to the shift in r. We want to prove that the shaded area is the relevant area for the measure of benefit and is identical to (4) or (5). To prove this, notice that the shaded area equals[14]

$$\int_{r_0}^{r_1}\int_{s_0}^{\infty} \frac{\partial}{\partial r}\left[Q\left(\frac{\partial C}{\partial s}\right)\right] ds\,dr$$
$$= \int_{s_0}^{\infty} Q(u, C(r_1, s), p_2)\frac{\partial C}{\partial s}(r_1, s)\,ds$$
$$- \int_{s_0}^{\infty} Q(u, C(r_0, s), p_2)\frac{\partial C}{\partial s}(r_0, s)\,ds$$

Now using the relation $C(r, s) = p_1$ we can rewrite the above expression as

$$(6) \quad \int_{C(r_1,s_0)}^{C(r_1,\infty)} Q(u, p_1, p_2)\,dp_1 - \int_{C(r_0,s_0)}^{C(r_0,\infty)} Q(u, p_1, p_2)\,dp_1 = -\int_{C(r_0,s_0)}^{C(r_1,s_0)} Q(u, p_1, p_2)\,dp_1$$

[11]With no rents, this banana is a rectangle.

[12]A Cobb-Douglas and a constant elasticity of substitution production function with a substitution elasticity below one satisfy this criterion.

[13]We use Shepherd's Lemma that the demand curve for input 1 equals $Q\partial C(r, s)/\partial r$. For simplicity I am assuming here that inputs 1 and 2 are used only in the production of output 1, and that the factors are in perfectly elastic supply.

[14]Again we are making use of Shepherd's Lemma that the demand curve for input 2 is $Q\partial C(r, s)/\partial s$.

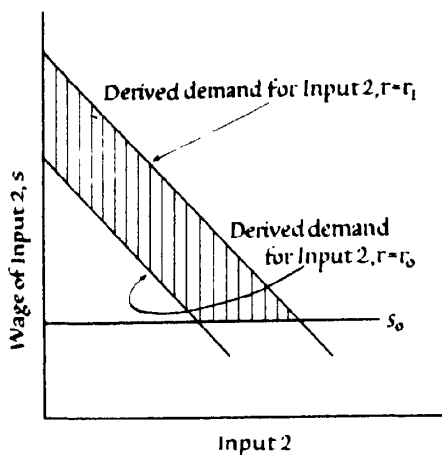



FIGURE 5

since by assumption[15]

$$\lim_{s\to\infty} \int_{C(r_0,s)}^{C(r_1,s)} Q(u, p_1, p_2)\,dp = 0$$

But (6) is identical to (5) and therefore to (4), hence the result is proved.

Figure 6 illustrates the three equivalent measures of gain in the output market, directly affected input (input 1) market, and in a related input (input 2) market. As before, the ability to use related input-market measures should be of use when data on directly affected markets are unavailable.[16]

[15]If this condition does not hold, or if the values of $C(r_0, \infty)$ and $C(r_1, \infty)$ are finite and unequal but known, then we would have to subtract from (6) the expression

$$\int_{C(r_0,\infty)}^{C(r_1,\infty)} Q(u, p_1, p_2)\,dp_1$$

If we further assume that we know an s_0^* and s_1^* such that $C(\infty, s_0^*) = C(r_0, \infty)$ and $C(\infty, s_1^*) = C(r_1, \infty)$, then this last expression can be written as

$$\int_{s_0^*}^{s_1^*} Q\frac{\partial C}{\partial s}(\infty, s)\,ds$$

which is a calculable area under the demand curve for input 2.

[16]The case when factors earn rents is a straightforward generalization of the proofs of Section I. When factors do earn rents, an interesting question is whether the rent change in a related factor market corresponds to the benefits that result from a technological improvement in some other factor. For example, will the change in agricultural rents reflect the benefits of the introduction of a new harvesting technology? It is possible to show (the proof is available on request) that when a related factor is in perfectly inelastic supply and when the technology is fixed coefficients, then the rent change in a related factor market will reflect the benefits of a technological change in some other factor. This last case appears to be the only one in which the relation between rent change and benefit measure holds in general.

III

The preceding discussion has shown how prior information on observable input markets can reveal the desired information about an unobservable input market. I now give an example where prior information on observable output markets yields the desired information about an unobservable output market. The example thus shows how information about other output markets can be used to measure benefits in an unobservable output market.

Suppose that we wish to evaluate the benefit of a technological improvement that lowers the price of the nth commodity from q_0 to q_1. Assume that the price vector p of commodities 1 through $n-1$ are unchanged. The analyst knows the (compensated) demand curve for commodities $1, \ldots, n-1$. Information on demand for commodity n is unavailable. If $m(u, p, q)$ represents the expenditure required to achieve utility level u at prices p, q, then we wish to calculate $m(u, p_0, q_0) - m(u, p_0, q_1)$ or

$$\int_{q_0}^{q_1} \frac{\partial m}{\partial q}\,dq$$

which is just the area under a compensated demand curve for commodity n between prices q_0 and q_1 (the derivative of the expenditure function with respect to price yields the compensated demand curve). Since no information on the compensated demand for commodity n is available, the calculation cannot be done.

Now let the analyst be given just a little bit of information about the demand for commodity n. Suppose he knows that at prices p^* for commodities $1, \ldots, n-1$, the quantity demanded of commodity n is zero at prices q_0 and q_1. For example, the demand for

FIGURE 6

tennis rackets can be driven to zero by sufficiently lowering the price of substitutes (golf clubs, squash rackets, etc.) and sufficiently raising the price of complements (tennis balls, sneakers, tennis court fees). The implication of the analyst's information is that $m(u, p^*, q_0) = m(u, p^*, q_1)$ or equivalently that the price of commodity n does not influence consumer expenditure when commodity n is not purchased. Our analyst can use his little bit of information to perform the desired measurement.[17]

Using the definition of a line integral and the fact that $m(u, p^*, q_0) = m(u, p^*, q_1)$ we can write that

$$(7) \quad m(u, p_0, q_0) - m(u, p_0, q_1) = [m(u, p_0, q_0) - m(u, p^*, q_0)] + [m(u, p^*, q_1) - m(u, p_0, q_1)]$$

$$= \oint_{p^*}^{p_0} \frac{\partial m}{\partial p}(u, p, q_0)\,dp + \oint_{p_0}^{p^*} \frac{\partial m}{\partial p}(u, p, q_1)\,dp$$

By assumption, the analyst knows the compensated demand curves $(\partial m/\partial p)$ for commodities $1, \ldots, n-1$. Therfore both integrals in (7) are calculable since they are just integrals under these $n-1$ compensated demand curves. Moreover, it follows that if the analyst has enough prior information about p^* so that he can calculate (7) for any q_0, q_1, then the analyst can let $q_1 = q_0 + \epsilon$ for small ϵ and thereby calculate $\partial m/\partial q$ which is precisely the compensated demand curve for commodity n.

[17]This same idea is discussed in detail by Karl-Göran Mäler, David Bradford and Gregory Hildebrand, and A. Myrick Freeman in the context of identifying demands for public goods.

IV

Missing data hamper many practical applications of cost-benefit analysis. This paper has shown that even when data on directly affected markets are unavailable, it is often still possible to perform the correct calculations by using available data from related output or other input markets. Such "indirect" calculations of costs and benefits should be a powerful and useful tool in applied cost-benefit analysis.

REFERENCES

J. Anderson, "The Social Cost of Input Distortions: A Comment and A Generalization," *Amer. Econ. Rev.*, Mar. 1976, *66*, 235–38.

D. Bradford and G. Hildebrandt, "Observable Preferences for Public Goods," *J. Publ. Econ.*, Oct. 1977, *8*, 111–31.

P. Diamond and D. McFadden. "Some Uses of the Expenditure Function in Public Finance," *J. Publ. Econ.*, Feb. 1974, *3*, 3–21.

A. M. Freeman. "The Theoretical Basis for Empirical Measurements of Benefits," mimeo., Resources for the Future, 1977.

A. C. Harberger, "The Measurement of

Waste," *Amer. Econ. Rev. Proc.*, May 1964, *54*, 58–76.

Kare-Göran Mäler, *Environmental Economics: A Theoretical Inquiry*, Baltimore 1974.

J. Panzar and R. Willig. "On the Comparative Statics of a Competitive Industry with Inframarginal Firms," *Amer. Econ. Rev.*, June 1977, *68*, 474–78.

R. Schmalensee, "Another Look at the Social Valuation of Input Price Changes," *Amer. Econ. Rev.*, Mar. 1976, *66*, 239–43.

D. Wisecarver, "The Social Cost of Input-Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–71.

Uncertainty and the Bidding for Incentive Contracts

By CHARLES A. HOLT, JR.*

In many procurement contracts, the monetary reward for the winning bidder is a function of the bid and of the *ex post* production cost. Therefore, contract incentives can affect the strategic relationship between a firm's bid and its estimates of production expenses. In the existing literature on incentive contract bidding, the effects of contract incentives on a firm's bidding strategy are usually analyzed with the assumption that expectations about rival bids are exogenous. That is, these expectations are summarized by a probability distribution that is assumed to be unchanged as incentives and other structural parameters are varied. But any factor that would alter one bidder's strategy would presumably alter rivals' bidding strategies as well, thereby changing the observed distribution of rival bids.

An alternative approach used in this paper is to require that bidders' expectations satisfy an equilibrium or rationality condition. A tractable model with endogenously determined expectations is possible when all bidders have similar preferences and opportunities. This model is outlined in Section I, and the Nash equilibrium bidding strategies of all firms are characterized in Section II. The focus in Section III is on the equilibrium effects of changes in incentives and other structural parameters on bidding behavior and expected procurement costs.

I. The Model

The awarding of contracts on the basis of competitive bidding is a frequently used resource allocation mechanism in both the public and private sectors. Consider a model of this competition in which the buyer announces a contract for the procurement of a specified commodity. It is assumed that bids are solicited and the contract is awarded to the lowest bidder. This model is only a first approximation of many procurement processes. For example, John Cross observed that sometimes the U.S. Department of Defense will propose a "rough contract," invite bids, and negotiate the final details with the winner. Often, a government agency or other buyer will engage in competitive negotiations with several bidders simultaneously.[1] Negotiations may be focused on price, but other considerations such as design and contract incentives may be involved. These nonprice dimensions will be ignored in this paper, and therefore, the propositions derived will be more relevant to situations in which price competition is important.

It is assumed that there are N bidders for a particular contract, and that the firm submitting the lowest bid is awarded the contract. In general, the contract specifies this firm's total receipts from the buyer as a function of the bid p and of other variables such as the *ex post* cost of production c. For the case of a linear incentive contract, the firm's total contract revenue is determined:

$$(1) \qquad \alpha p + \beta(p - c) + c$$

where $0 \leq \alpha \leq 1$ and $0 \leq \beta \leq 1$. If $\beta = 0$, the firm receives the production cost plus a "fixed fee" αp. A fixed price contract is one for which $\alpha = 0$ and $\beta = 1$; the winning firm's profit equals the full difference between the bid price and the production cost. Note that if the cost exceeds the bid price, this "overrun" is penalized by a factor of β. Underruns are

*Department of economics, University of Minnesota. I wish to thank Morris DeGroot, Edward Prescott, Michael Visscher, and M. Thérèse Flaherty for guidance and advice. I am also indebted to an anonymous referee whose suggestions improved the paper.

[1]It is not uncommon for buyers to negotiate contract terms with a single firm. However, government agencies are often required by law to seek competitive offers unless only one firm is capable of providing the desired commodity.

rewarded symmetrically. The parameters α and β are sometimes referred to as the "guaranteed profit rate" and the "sharing rate," respectively.

Many contracts that are awarded on the basis of competitive bidding are for tailor-made jobs that cannot be readily purchased in the existing markets. This is one of the reasons why there may be considerable uncertainty about the cost of completing the contract job. Note that a reduction in the sharing rate β means that the government or other procuring agent agrees to accept both a greater share of the risk of a cost overrun and a greater share of the potential savings resulting from an underrun. A common justification for the use of incentive contracts is that if bidders are risk averse, then the procuring agent may be able to reduce its expected procurement cost by accepting a greater share of the risk. The assumption is that firms bidding for contracts cannot insure against the risk of cost overruns because of moral hazard problems: overruns are sensitive to the firm's own decisions. If the owners of a firm are risk averse, the management may be reluctant to undertake a relatively large contract with significant cost uncertainties that cannot be insured. In addition, the managers themselves may be risk averse.[2]

First, consider the bidding decision of a specific firm. It is assumed that this firm's preferences toward profit risks can be represented by a von Neumann-Morgenstern utility function $U(\cdot)$.[3] It is also assumed that the firm does not know its production cost with certainty at the time of bidding. Let the expected value of the production cost be denoted $\bar{c}$. The actual *ex post* production cost c is determined:

$$c = \bar{c} + \gamma \tag{2}$$

where γ is the realization of a continuous random variable with mean 0. The probability density function (*PDF*) and distribution function (*DF*) of γ will be denoted by $h(\cdot)$ and $H(\cdot)$, respectively. It is assumed that $h(\cdot)$ is strictly positive on an open interval (c_1, c_2).

Suppose that the firm under consideration is the ith firm and that its bid is p_i. It is apparent from (1) that this firm's contract profit, as a function of both the bid p_i and the realized cost c, is $\alpha p_i + \beta(p_i - c)$. Therefore, its expected profit, denoted by $\bar{z}$, is $(\alpha + \beta)p_i - \beta\bar{c}$. The firm's expected utility of the contract profit, denoted by $E_c(p_i)$ is

$$E_c(p_i) = \int_{c_1}^{c_2} U(\bar{z} - \beta\gamma)\, h(\gamma)\, d\gamma \tag{3}$$

Following John McCall's model of contract bidding, firms that lose the contract are assumed to use their resources in the private sector. The ith firm's profit from the alternative private sector operation will be denoted r_i. The alternative profit r_i is assumed to be known to firm i at the time the bid is submitted. Therefore, the firm's utility is $U(r_i)$ in the event that the contract is lost. If θ is the probability of winning the contract, then the firm's expected utility can be written

$$\theta E_c(p_i) + [1 - \theta]U(r_i) \tag{4}$$

In general, θ will be a function of the bid p_i: high bids will be less likely to win. The implicit assumption in David Baron's analysis of incentive contract bidding is that the functional relationship between θ and p_i is independent of the structural parameters of the model. But the relation between p_i and θ surely depends on the bidding strategies of rival firms. For example, consider a change in contract incentives that would cause a change in the ith firm's optimal bid for a given θ function. It seems reasonable to expect this type of change to alter rival bids as well, and therefore the ith firm's probability of winning with a bid p_i should be altered.

In order to determine the *equilibrium* effects of structural changes on bidding behavior, it is necessary to analyze the bidding decisions of all bidders simultaneously. The model is manageable if all bidders are assumed to have similar preferences and opportunities. Specifically, it is assumed that all N firms have the same utility function for

[2] On the basis of case studies, F. M. Scherer, p. 275, reported a prevalent concern that short-run contract losses might be interpreted as managerial failure requiring reorganization.

[3] The justification of a utility function for the firm should be based on the preferences of managers and owners, and on the implicit contracts made by these agents under conditions of uncertainty. In my opinion, this is an important unsolved problem.

profit. This common utility function $U(\cdot)$ is twice differentiable, strictly increasing, and strictly concave. In addition, all firms are assumed to be equally efficient in the sense that the probability distribution of the cost c is the same for each firm.[4] Thus the ith firm's expected utility as a function of p_i and r_i is given in (4) for firms $i = 1, \ldots, N$.

Although a firm's bid is often referred to as a "target cost," these bids will be analyzed as strategic moves, not as forecasts or goals. A bidding strategy specifies a relationship between the bid price and the firm-specific parameter r_i. This alternative profit opportunity may differ from firm to firm at any particular time because of differences in planned investment projects. However, firms have similar opportunities in the sense that the profit alternatives are assumed to be independent realizations of a continuous random variable with a *PDF* and *DF*, denoted by $g(\cdot)$ and $G(\cdot)$, respectively. It is assumed that $g(\cdot)$ is strictly positive on a finite interval (a, b).[5] This density is known by all so that firms have symmetric information about the profit opportunities of rivals.[6]

To summarize the model, all players or firms know that all utility functions and cost distributions are the same. Each firm knows its own alternative profit opportunity, and firms have symmetric information about rivals' opportunities. The contract parameters are announced, sealed bids are submitted, and the contract is awarded to the lowest bidder.

II. Equilibrium Bidding Strategies

The bidding competition will be analyzed as a noncooperative game.[7] Because of the symmetry of information and preferences, only symmetric Nash equilibria will be considered. In a symmetric equilibrium the common Nash strategy will have the property that it maximizes the expected utility of each bidder when all rivals are known to be using the same strategy. Recall that a bidding strategy is a function of the alternative profit opportunity. A Nash equilibrium strategy, together with the distribution of rival profit alternatives, determines the distribution of rival bids. Consequently, the equilibrium strategy maximizes the firm's expected utility with respect to this distribution of rival bids. If all firms use the Nash strategy, each firm's subjective probability distribution over rival bids corresponds to the distribution of rival bids prior to the drawing of profit alternatives. A Nash point in strategies generates a set of "rational expectations" in the sense of Robert Lucas and Edward Prescott. In equilibrium, expectations are endogenously determined by assuming that each firm uses a Nash bidding strategy.

Suppose that the common equilibrium bidding strategy is a strictly increasing function $p = p(r)$ defined for $r \in (a, b)$. Then the firm with the lowest alternative profit opportunity would win the contract. This firm i would win if

$$r_i < \min_{j \neq i} \{r_j\}$$

Let y denote the minimum rival profit opportunity, i.e., y is lowest order statistic from a sample of size $N - 1$ from a distribution with *DF* $G(\cdot)$. Let the *DF* and *PDF* of y be denoted by $F(y)$ and $f(y)$, respectively; these are computed:

$$(5) \qquad F(y) = 1 - [1 - G(y)]^{N-1}$$
$$f(y) = dF(y)/dy$$

A symmetric argument shows that this $F(y)$ would represent each firm's probabilistic beliefs about the minimum of its rivals' profit opportunities.[8] Thus the functions $F(y)$

[4]The critical nature of this assumption is discussed in Section III.

[5]It is straightforward to generalize the analysis to cover the cases in which a and b are not finite. See my dissertation, p. 85.

[6]Note that the only relevant difference between firms is the r_i parameter, because the $U(\cdot)$ function and the density functions $h(\cdot)$ and $g(\cdot)$ are the same for each firm. If any of these symmetry assumptions were relaxed, the bidding strategies would depend on all firm-specific parameters.

[7]The game-theoretic approach taken in this paper is based on William Vickrey's initial analysis of bidding games, and on the subsequent work of Armando Ortega-Reichert and Robert Wilson.

[8]An anonymous referee pointed out that the qualitative implications of this bidding model are independent of the properties of $G(\cdot)$. It is important that the distribution function $F(\cdot)$ be the same for all bidders. The assumption that the profit opportunities are independent realizations

and $p(r)$ would determine each firm's beliefs about its lowest rival bid $p(y)$. These beliefs, in turn, would determine a firm's probability of winning with a specific bid; this is the probability θ in the firm's objective function (4).

This connection between bidding strategies and the probabilities of winning is presented in the Appendix. There it is shown that there is a symmetric equilibrium bidding strategy $p(r)$ that is implicitly determined:

$$E_c(p(r)) = \frac{\int_r^b U(s) f(s)\, ds}{1 - F(r)} \qquad r \in (a, b) \tag{6}$$

Recall that $E_c(p)$ defined in (3) is the expected utility of contract profit if a bid p is accepted. The expression on the right-hand side of (6) can be interpreted as being the expected utility of the minimum rival profit opportunity y, conditional on the event that $y > r$. Then (6) states that this conditional expected utility equals the expected utility of contract profit evaluated at the equilibrium bid.

Without an assumption of risk neutrality or constant absolute risk aversion, it is impossible to obtain a closed-form expression for the equilibrium bidding strategy from equation (6). However, the qualitative effects of changes in contract parameters on bids and contract profits can be determined regardless of the assumed degree of risk aversion.

First, it is convenient to introduce some new notation. Let $\phi(\cdot)$ and $\Phi(\cdot)$ denote the respective *PDF* and *DF* of the random variable $\epsilon = \beta\gamma$, so that, for $\beta > 0$,

$$\Phi(\epsilon) = H(\epsilon/\beta), \text{ and } \phi(\epsilon) = \frac{1}{\beta} h(\epsilon/\beta) \tag{7}$$

Then it follows from (3) and the definition of ϵ that

$$E_c(p) = \int_{c_1}^{c_2} U(\bar{z} - \epsilon)\phi(\epsilon)\, d\epsilon = U(\bar{z} - \Delta_\epsilon) \tag{8}$$

of the same random variable is one way of motivating the symmetric information implied by $F(\cdot)$.

where Δ_ϵ is a nonnegative risk premium.[9] With this notation, equation (6) which determines the equilibrium bid can be written

$$U(\bar{z} - \Delta_\epsilon) = U^*(r) \tag{9}$$

where

$$U^*(r) = \frac{\int_r^b U(s) f(s)\, ds}{1 - F(r)} \tag{10}$$

For each value of $r \in (a, b)$, it can be shown that equation (9) uniquely determines the expected contract profit $\bar{z}$ that corresponds to the equilibrium bid $p(r)$. First note that $U(\bar{z} - \Delta_\epsilon)$ defined in (8) is a continuous, strictly increasing function of $\bar{z}$ as shown in Figure 1. The unique equilibrium value of $\bar{z}$ that solves (9) is determined by the intersection of the $U(\bar{z} - \Delta_\epsilon)$ line with the horizontal $U^*(r)$ line. Also, it is apparent from (10) that $U(r) < U^*(r) < U(b)$. If $\bar{z} = r$, then $U(\bar{z} - \Delta_\epsilon) \leq U(r) < U^*(r)$ and it follows that the equilibrium solution for $\bar{z}$ is greater than r. For each value of r, the equilibrium level of $\bar{z}$ will be denoted by $\bar{z}(r)$.

The Nash bid for each value of r is determined from the equilibrium level of $\bar{z}(r)$:

$$p(r) = \frac{\beta\bar{c} + \bar{z}(r)}{\alpha + \beta} \tag{11}$$

If the utility function exhibits constant absolute risk aversion, then the risk premium Δ_ϵ defined in (8) will be a constant, independent of $\bar{z}$.[10] In this special case, it follows from (9) that $\bar{z} = \Delta_\epsilon + U^{-1}(U^*(r))$, where $U^{-1}(\cdot)$ denotes the inverse of $U(\cdot)$. The equilibrium bidding strategy can be determined explicitly from (11) in this case.

It can be shown that the bidding strategy determined by (9) and (11) is unique in the class of symmetric, differentiable strategy equilibria.[11] Finally, it can be shown that the

[9] A risk premium is defined by John W. Pratt to be the quantity Δ that satisfies: $E\{U(x)\} = U(E\{x\} - \Delta)$, where the expectation is taken with respect to the distribution of the random variable x.

[10] See Pratt for a discussion of constant absolute risk aversion and risk premiums.

[11] It follows from (9), (10), and the differentiability of $U(\cdot)$ that the equilibrium strategy $p(r)$ is a strictly

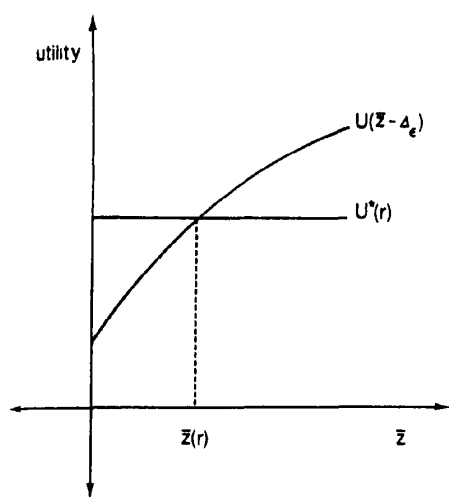



FIGURE 1

equilibrium $p(r)$ bids are indeed Nash equilibrium bids; no firm can increase its expected utility by submitting a bid $p \neq p(r)$ when all $N-1$ rivals are known to use the equilibrium strategy function implicit in (9) and (11).[12]

III. The Equilibrium Effects of Structural Changes

A structural change is an event that alters the decision environment of all potential bidders. The bidding behavior of all competitors is generally affected by a structural change, and therefore, each bidder's expectations concerning rival bids should be affected.

First, consider the contract parameters α and β. Note that $U^*(r)$ defined in (10) is obviously independent of α and β. Therefore, it follows from (9) that the expected utility of the contract profit (evaluated at the equilibrium bid) is independent of the contract parameters.

However, an increase in β will increase equilibrium bids. To see this, consider the effect of a change in β on the distribution of ϵ. It follows from (7) that an increase from β_1 to β_2 will change the DF of ϵ from $\Phi_1(\epsilon) = H(\epsilon/\beta_1)$ to $\Phi_2(\epsilon) = H(\epsilon/\beta_2)$. The mean of the random variable ϵ is 0 for each value of β, but $\Phi_2(\epsilon) > \Phi_1(\epsilon)$ if $\epsilon < 0$, $\Phi_2(\epsilon) = \Phi_1(\epsilon)$ if $\epsilon = 0$, and $\Phi_2(\epsilon) < \Phi_1(\epsilon)$ if $\epsilon > 0$. Next, it can be shown that $\Phi_1(\epsilon)$ is strictly greater than $\Phi_2(\epsilon)$ in the sense of second degree stochastic dominance, i.e., that $\int_{-\infty}^{\epsilon} [\Phi_2(x) - \Phi_1(x)]\, dx \geq 0$ for all ϵ and $>$ for at least one value of ϵ.[13] Given the assumption that $U''(\cdot) < 0$, it follows from a strict version of the second degree stochastic dominance theorem that the expected value of $U(\bar{z} - \beta_1\gamma)$ is strictly greater than the expected value of $U(\bar{z} - \beta_2\gamma)$ for each fixed value of $\bar{z}$ (see Vijay Bawa, Theorem 2, pp. 101–02). Equivalently, an increase in β, with $\bar{z}$ fixed, will reduce $U(\bar{z} - \Delta_\epsilon)$. This downward shift in the $U(\bar{z} - \Delta_\epsilon)$ curve in Figure 1 will increase the equilibrium level of $\bar{z}(r)$, and therefore, the equilibrium bids determined by (11) will increase. With all firms bidding for greater contract profits, the expected procurement cost will increase.

The conclusion is that in this model the government or other procuring agent could reduce its expected procurement cost by accepting all risk ($\beta = 0$).[14] This risk effect is decisive because there are no cost reduction activities in this model. An increase in β increases the penalty for cost overruns, and therefore, high values of β might induce firms to keep costs down once the contract has been

increasing, differentiable function of r. It follows from L'Hospital's Rule that the terminal condition implicit in (6) is

$$\lim_{r \to b} E_c(p(r)) = U(b)$$

In my dissertation, p. 90, I have shown that any symmetric equilibrium strategy must satisfy this condition, and that the uniqueness of the equilibrium strategy follows from the uniqueness of the terminal condition.

[12]A proof can be found in my dissertation, p. 90.

[13]This second degree stochastic dominance inequality obviously holds for $\epsilon \leq 0$. If $\epsilon > 0$,

$$\int_{-\infty}^{\epsilon} (\Phi_2(x) - \Phi_1(x)]\, dx =$$
$$\int_{-\infty}^{\infty} [\Phi_2(x) - \Phi_1(x)]\, dx - \int_{\epsilon}^{\infty} [\Phi_2(x) - \Phi_1(x)]\, dx$$

The first integral on the right-hand side of this equation is zero because the means of each DF are equal. Recall that $\Phi_2(\epsilon) < \Phi_1(\epsilon)$ for $\epsilon > 0$, and therefore, the second degree stochastic dominance inequality holds for $\epsilon > 0$ as well.

[14]However, a risk-averse procuring agent may not be willing to accept all risk.

awarded. A bidding model with cost reduction opportunities is analyzed in my dissertation, pp. 102–12.[15] High values of β result in lower expected production costs in that model. The value of β that minimizes expected procurement costs is shown to depend on the magnitude of this efficiency effect relative to that of the risk effect described above.

Cross has argued that efficiency is institutionalized; new techniques are implemented on a tentative basis, and they are retained if the results are satisfactory. This type of experimentation may not be possible or profitable on a short-run basis for a single contract. However, incentives may be more effective in the case of the procurement of standard equipment on a continuing basis. The risk effect may also depend on the type of project; the initial production of tailor-made projects may involve much more uncertainty than the continued production of standard equipment. This suggests the attractiveness of a flexible policy; cost-plus contracts (or contracts with a low β) could be used for initial procurements when the risk effect is likely to be significant. Then contracts with higher incentive rates could be used for subsequent purchases.

The effects of changes in β can be summarized:

PROPOSITION 1: *If 1) firms are risk averse, 2) production costs are uncertain, and 3) there are no cost reduction opportunities, then an increase in the sharing rate β will increase equilibrium bids. In addition, the expected value of the contract profit evaluated at each firm's equilibrium bid will increase. The expected procurement cost is therefore increasing in β.*

Next consider a shift in the α parameter. Recall that $\epsilon = \beta\gamma$, and therefore, the distribution of ϵ will not be altered. Thus for a fixed level of $\bar{z}$, both $U(\bar{z} - \Delta_\epsilon)$ in (8) and $U^*(r)$ in (10) are unaffected by the shift in α. It now follows from (9) that there is no change in $\bar{z}(r)$, the expected value of the contract profit evaluated at each firm's equilibrium bid. Conseqently, the expected procurement cost is independent of α. But an increase in α will decrease the equilibrium bid for each value of r because $\bar{z}(r)$ is unchanged and $\bar{z}(r) = (\alpha + \beta)\, p(r) - \beta\bar{c}$. Note that an increase in α that reduces equilibrium bids will increase the probability that there will be a cost overrun. To summarize:

PROPOSITION 2: *An increase in the guaranteed profit rate α will decrease equilibrium bids and increase the probability of cost overruns. However, the resulting overruns do not indicate excessive procurement expenditures because expected procurement costs are independent of α in this model.*[16]

Next, consider the effect of an increase in the number of bidders N. Recall that $U^*(r)$ in (10) is the expected utility of the minimum rival profit opportunity y, conditional on the event that $y \geq r$. As the number of rival bidders increases, the minimum rival profit opportunity is likely to be reduced. Therefore, one would expect that $U^*(r)$ is a decreasing function of N for each value of r. A proof of this conjecture is given in my dissertation, pp. 62–66. Thus, the increase in N shifts the $U^*(r)$ line downward in Figure 1, and therefore $\bar{z}(r)$ decreases. It follows that an increase in the number of bidders reduces equilibrium bids and expected procurement costs.

As the number of bidders approaches infinity, one would expect that if y is greater than r, it is not likely to be very much greater. Recalling the previous interpretation of $U^*(r)$ as a conditional expected utility, one would expect that $\lim_{N\to\infty} U^*(r) = U(r)$. This limit is verified in my dissertation, pp. 69–71. In the limit, (9) becomes $U(\bar{z} - \Delta_\epsilon) = U(r)$.

In order to interpret this limiting strategy, note that $U^*(r)$ in (10) exceeds $U(r)$. For *finite* N, the equilibrium strategy implicit in (9) has the property that $U(\bar{z}(r) - \Delta_\epsilon) >$

[15]Both Scherer and J. Michael Cummins discuss this cost reduction effect in the context of a negotiated contract.

[16]Cummins, p. 179, reaches the same conclusion in his analysis of negotiated contracts, that is, that cost overruns do not necessarily indicate excessive procurement costs.

$U(r)$. In other words, it is optimal to submit a bid for which the expected utility of the contract profit exceeds the utility of the alternative profit in the event of a loss. If this were not the case, there would be no reason to bid! However, as the number of bidders approaches infinity, the competition induces each bidder to bid for a contract profit that has the same expected utility as the alternative profit. This limiting strategy is the type of strategy used by McCall in an important analysis of incentive contract bidding when firms have different production costs.

Finally, it is interesting to consider the interrelationships between firms' aversions to risk and the riskiness of production costs. There are several plausible measures of the riskiness of a random variable, but the comparisons in this section are between distributions with identical means. A mean preserving increase in risk for a random variable $\tilde{x}$ is defined to be the creation of a new random variable $\tilde{x} + \tilde{\eta}$, where $\tilde{\eta}$ is independent of $\tilde{x}$ and $E\{\tilde{\eta}\} = 0$. Michael Rothschild and Joseph Stiglitz show that a mean-preserving increase in risk will reduce the expected utility of a risk averter. Equivalently, the risk premium is increased.

Consider a structural change that results in a mean-preserving increase in the risk associated with production cost for all bidders. As a result, the risk premium Δ_c determined in (8) will increase. Thus the $U(\bar{z} - \Delta_c)$ line in Figure 1 shifts downward for each value of $\bar{z}$, and the equilibrium level of $\bar{z}(r)$ is increased. In a Nash equilibrium, bids are increased because bidders demand a greater expected contract profit as compensation for the increased risk.[17] To summarize:

PROPOSITION 3: *If bidders are risk averse, then a mean-preserving increase in risk associated with each firm's production cost will increase both the equilibrium bid $p(r)$ and the equilibrium contract profit $\bar{z}(r)$ for each value of r.*

A major restrictive assumption in this paper's symmetric model is that the probability distribution of production costs is the same for all firms. If some firms were in fact more efficient on the average, than it would be important to design contracts that give efficient firms an advantage in the bidding competition. However, the necessary asymmetry will make the analysis of equilibrium bidding quite difficult in such a model.

IV. Conclusion

In equilibrium, the structural factors that form the decision environment of all bidders will have an impact on the probability distribution of bids relevant for each firm's decision. Consequently, it is not meaningful to consider the effect of a structural change on a firm's bidding behavior when expectations about rival bids are assumed to be unchanged. Expectations about rival bids were assumed to be exogenous in Baron's analysis, and the effects of contract incentives on bidding behavior were generally indeterminate. This is because the probability distribution representing a firm's expectations was essentially arbitrary; no equilibrium condition was imposed on these expectations.

For the symmetric model analyzed in this paper, firms' expectations about rivals bids are endogenous and rational when firms use their Nash equilibrium bidding strategies. The equilibrium effects of contract incentives and other structural parameters on both bidding strategies and expected procurement costs are determined. Briefly, the effect of an increase in the "guaranteed" profit rate is to lower bids to the extent that each firm's expected contract profit for the bid tendered is unchanged. Thus a change in the profit guarantee will not affect expected procurement costs, but it is shown that the frequency of cost overruns may be affected. Consequently, procurement efficiency is not necessarily associated with the absence of observed

[17] Baron analyzes the effect of cost uncertainty on a particular firm's bid when the firm's probability distribution of the minimum rival bid is fixed and independent of the firm's production cost uncertainty. Baron shows that the optimal bid for a deterministic production cost $\bar{c}$ is less than the optimal bid if costs are uncertain with an expected value of $\bar{c}$. The proof of this proposition is based on an assumption that the firm's utility function exhibits nonincreasing absolute risk aversion.

cost overruns. On the other hand, an increase in the sharing rate will induce risk-averse firms to bid for greater expected contract profits as compensation for accepting a larger share of the risk associated with uncertain production costs. The main results were stated more precisely in Propositions 1–3.

APPENDIX

There are two steps in the analysis of equilibrium bidding strategies. First, it is useful to consider the optimal bid of a typical firm i when its $N-1$ rivals are known to use a strategy $p_j = p(r_j), j \neq i$. The next step is to determine the $p(\cdot)$ function that is the firm's optimal strategy when all rivals are known to be using the *same* strategy function; we are looking for a fixed point in strategy functions.

Suppose that the strategy function used by the ith firm's $N-1$ rivals is a function $p(\cdot)$ which is differentiable and strictly increasing on (a, b). Then there is an inverse strategy $\pi(\cdot)$ such that $\pi(p(r_j)) = r_j$ and $\pi'(\cdot) > 0$. Recall that

$$y = \min_{j \neq i} \{r_j\}$$

so it follows from the assumed monotonicity of $p(\cdot)$ that the lowest rival bid is $p(y)$. Thus the ith firm will win the contract if $p_i < p(y)$, or equivalently, if $\pi(p_i) < y$. The DF of y is denoted by $F(\cdot)$, so the probability of winning a contract with a bid of p_i is $1 - F(\pi(p_i))$. Therefore, the ith firm's expected utility is given in (4) with $\theta = 1 - F(\pi(p_i))$. The necessary condition for determining the optimal bid p_i is

$$\text{(A1)} \quad E_c'(p_i)[1 - F(\pi(p_i))] - [E_c(p_i) - U(r_i)] f(\pi(p_i))\pi'(p_i) = 0$$

where $E_c'(p_i) = dE_c(p_i)/dp_i$. For a given rival strategy $p(\cdot)$ corresponding to $\pi(\cdot)$, equation (A1) specifies a bid p_i for any $r_i \in (a, b)$.

In a symmetric Nash equilibrium, all firms will use the same strategy function: $p_i = p(r_i)$ and $r_i = \pi(p_i)$ for $i = 1, \ldots, N$. If one replaces r_i in (A1) with $\pi(p_i)$, the result is a differentiatial equation in the equilibrium $\pi(\cdot)$ function. Equivalently, the equilibrium bidding strategy function $p(\cdot)$ must satisfy the differential equation:

$$\text{(A2)} \quad [1 - F(r_i)]E_c'(p(r_i))p'(r_i) - [E_c(p(r_i)) - U(r_i)] f(r_i) = 0$$

because $1/\pi'(p_i) = p'(\pi(p_i)) = p'(r_i)$. Equation (A2) must be satisfied for all r_i on the interval (a, b). Therefore, (A2) can be integrated from any arbitrary value of $r \in (a, b)$ to b, and the integrated equation is

$$\text{(A3)} \quad \int_r^b [1 - F(s)] \frac{dE_c(p(s))}{ds} ds = \int_r^b [E_c(p(s)) - U(s)] f(s)\, ds$$

because $E_c'(p(r_i))\, p'(r_i) = dE_c(p(r_i))/dr_i$. Equation (6) in the text is obtained by integrating the left-hand side of (A3) by parts.

REFERENCES

D. P. Baron, "Incentive Contracts and Competitive Bidding," *Amer. Econ. Rev.*, June 1972, *62*, 384–94.

V. S. Bawa, "Optimal Rules for Ordering Uncertain Prospects," *J. Finan. Econ.*, Mar. 1975, *2*, 95–121.

J. G. Cross, "A Reappraisal of Cost Incentives in Defense Contracts," *Western Econ. J.*, June 1968, *6*, 205–25.

J. M. Cummins, "Incentive Contracting for National Defense: A Problem of Optimal Risk Sharing," *Bell J. Econ.*, Spring 1977, *8*, 168–85.

C. A. Holt, "Bidding for Contracts," in Arnold Zellner, ed., *Bayesian Analysis in Economic Theory and Time-Series Analysis—The 1977 Savage Dissertation Awards*, Amsterdam forthcoming.

R. E. Lucas and E. C. Prescott, "Investment Under Uncertainty," *Ecomometrica*, Sept. 1971, *39*, 659–81.

J. J. McCall, "The Simple Economics of Incentive Contracting," *Amer. Econ. Rev.*, Dec. 1970, *60*, 837–46.

A. Ortega-Reichert, "Models for Competitive Bidding Under Uncertainty," tech. rept. no. 103, Dept. Operations Res., Stanford Univ., Jan. 1968.

J. W. Pratt, "Risk Aversion in the Small and in

the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

M. Rothschild and J. E. Stiglitz, "Increasing Risk: I. A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43.

F. M. Scherer, "The Theory of Contractural Incentives for Cost Reduction," *Quart. J. Econ.*, May 1964, *78*, 257–80.

W. Vickrey, "Counterspeculation, Auctions, and Competitive Sealed Tenders," *J. Finance*, Mar. 1961, *16*, 8–37.

R. B. Wilson, "Competitive Bidding with Disparate Options," *Manage. Sci.*, Mar. 1969, *15*, 446–48.

Price Controls and Optimal Export Policies under Alternative Market Structures

By RICHARD E. JUST, ANDREW SCHMITZ, AND DAVID ZILBERMAN*

International trade is a transaction which takes place among institutions. Competitive producers rarely sell directly to competitive consumers in a world trading market! For example, a wheat producer in Kansas does not negotiate his wheat sales directly with Russian grain purchasers. It is common for private or semipublic institutions, such as international trading companies or marketing boards, to buy from producers and sell to consumers. In many exporting nations, marketing boards or organizations exist (for example, the Organization of Petroleum Exporting Countries, *OPEC*; Canadian and Australian Wheat Boards; Ghana Cocoa Board; Uganda Coffee Board) which are essentially producer boards that do the actual selling of the product both within and outside the country. In the United States, private grain traders who buy grain from thousands of *U.S.* farmers are the dominant sellers of *U.S.* grain domestically and to foreign markets.

The literature to date on optimal tariff policies does not treat explicitly the nature and role of firms engaged in international marketing. In the classical optimal tariff literature, there are essentially three sectors: competitive producers; competitive consumers; and the government. The latter uses its power to intervene in international trade to maximize domestic social welfare. Recently, alternative models have been constructed in which the government uses tariff policy to achieve other goals. For example, Harry Johnson (1951, 1968) and Trent Bertrand have developed the "optimum government revenue tariff." In their models the government plays the role of the marketing agency using its monopoly and monopsony power to maximize the revenue it collects from international trade.

In this paper an international trade model for an exporting country is constructed in which international trading firms may exercise the type of market power Johnson (1951) and Bertrand attributed to the government sector; the government plays the traditional role of using tariff and other policies to maximize domestic social welfare.[1] The model thus consists of four sectors: 1) competitive producers; 2) competitive consumers; 3) the government; and 4) international trading institutions which are assumed to be noncompetitive. The welfare effects on society from the activities of two types of noncompetitive trading institutions are analyzed separately, and appropriate governmental actions for achieving maximum social welfare are suggested.[2] Explicitly, a marketing firm with both monopoly and monoposony power is considered where the firm operates independently from producers and consumers to maximize its own profits. This is an extension of results relating to the kind of firm introduced in Abba Lerner's classic paper on the measurement of monopoly power. A second case deals with international transactions carried out by producer cartels. Their activities are discussed both with and without government controls.

It is important to recognize that governmental controls are often imposed on interna-

*Department of agricultural and resource economics, University of California-Berkeley. We wish to thank an anonymous referee for useful comments on an earlier version of this paper. Giannini Foundation Paper No. 513.

[1]The models used are partial in nature in that only exporting firms are considered. Partial analysis is used because of simplicity and sharpness of results. Paul Samuelson (1971) developed conditions under which partial analysis leads to the same results as does the general equilibrium approach.

[2]The welfare sections of this paper are based on the classic concept of economic surplus. The limitations of this type of analysis, which have been discussed by John Currie, John Murphy, and Schmitz, should be borne in mind.

tional trading institutions. As examples, for the first time since World War II, the United States imposed export controls on grain shipments and required that export sales to the Soviet Union by the private trading companies be reported if the sales were over a certain volume, and the Canadian government imposed a ceiling on the price that millers in Canada pay for wheat, while the price for export sales is set by the Canadian Wheat Board (which has among its objectives the maximization of producer returns) for export sales.[3] The specific control examined in this paper is somewhat similar. The government in the exporting country permits the marketing institution to charge whatever price it can abroad but, internally, the price it charges domestic consumers must equal the (marginal) "costs" of production. Interestingly, this type of government regulation does not lead to maximum welfare for the exporting country as a whole. To attain welfare-efficient trade, an export *tax* is needed along with this type of control in the case where the export firm is a cartel or marketing board acting in the interest of producers (as shown in Section II); and an export *subsidy* is needed in the case where the export firm is a private monopoly-monopsony marketing firm (as shown in Section III). The conclusion that a subsidy is needed in the latter case simply states that the implicit export tax used in the firm's profit-maximizing calculations exceeds the implicit export tax that should be used if the firm were operating competitively.[4]

The debate over whether or not multinational marketing companies are competitive or not has occupied a large part of the economic literature and antitrust court cases. The results are inconclusive. However, in view of the recent behavior of such organizations as *OPEC* and other marketing boards and the description of firms presented by Jim Hightower, it seems worthwhile to consider, at least theoretically, marketing firms which have market power. In addition, since the results from competition are well known, the analysis of alternative polar cases is of interest.

Readers familiar with the literature on market distortions (particularly factor-market imperfections) in the framework of international trade theory will recognize in this paper an application of the general principle that in order to achieve maximum welfare it is necessary to adopt a subsidy/tax scheme which offsets the domestic distortion, *plus* a tax aimed at taking advantage of the country's monopoly and/or monopsony power in trade. In this case an export tax is required for the latter purpose. Since a tax is equivalent to a subsidy to domestic consumers *and* a tax on domestic production, the optimum trade controls can be estimated by simply imposing taxes or subsidies on domestic concerns.

I. Optimal Trade for an Individual Country

In this section the optimal trade conditions for an individual country exporting a single good are reviewed. It can easily be shown that these conditions can be attained under competition by imposing an export tax. In subsequent sections, optimal policies under alternative noncompetitive institutional frameworks are determined in an analogous manner. In other words, controls are developed in each case which lead to the same pattern of trade as indicated by the optimal conditions derived in this section. This forms the basis for the optimization performed in succeeding sections.

Consider the two-country model in Figure 1: country A exports good x to country B. The domestic demand in country A is D_A, and B's demand for country A's exports is D_E.[5] The

[3]For a discussion of how the Canadian Wheat Board was established and the extent to which it serves producers' interests and enjoys producers' support, see Vernon Fowke.

[4]This result is similar to the proposition that maximum revenue (rather than maximum welfare) tariff exceeds the optimum tariff. On this, see Johnson (1951), Edward Tower, and W. M. Corden.

[5]If none of good x is produced in B, the demand schedule D_E is comparable to D_A in country A. However, if x is also produced in country B, the demand schedule D_E is an excess demand curve. In this case the results of the paper also hold since it is assumed that the marketing firm (either the board or the middleman) does not use its monopsony power against producers in B (i.e., the firm does not use the marginal outlay schedule to the supply schedule in B in calculating optimal conditions).

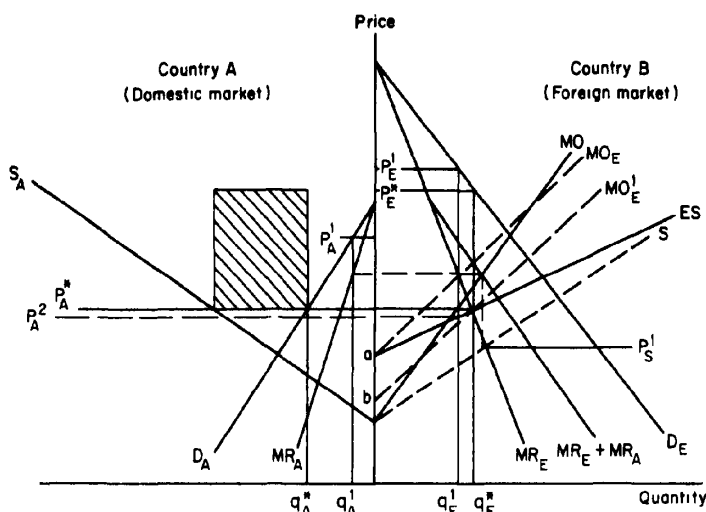



FIGURE 1. FIRM WITH MONOPOLISTIC AND MONOPSONISTIC POWER

supply function in A for the *competitive* industry producing x is S. Assume that A maximizes welfare from trade in x. Its welfare criterion is to maximize the sum of the social surpluses from consumption and exports minus the social cost of production. For competition in both production and demand, the exporting country must impose a tax or a quota on exports in order to achieve the optimum allocation of resources. The solution for the optimal export tax for country A can be determined by constructing an excess supply curve (ES) for country A and observing its intersection with the marginal revenue schedule MR_E which corresponds to demand D_E. In the optimal solution *the price to both consumers and producers in country A is the same,* P_A^*, and the price in country B is P_E^*.[6] The tax is given by $P_E^* - P_A^*$, and the tax revenue collected by country A on exports is represented by the crosshatched area. From the mathematical treatment in the Appendix, this tax can be calculated by $q_E^* D'(q_E^*)$ or, equivalently, $-P_E^*/\eta_E$, where η_E, is the elasticity of foreign excess demand.

[6]It is assumed that all demand curves are negatively sloped, all supply curves are positively sloped, and revenue functions are concave.

II. The Monopoly-Monopsony Model

Consider the monopoly-monopsony case where a single firm buys from competitive producers in country A and sells to competitive consumers in country B. Ignoring marketing costs, the optimization conditions for the monopoly-monopsony case are given in the Appendix and are illustrated geometrically in Figure 1.[7] To determine the optimum pricing policy that the marketing firm will use, a marginal outlay curve MO (i.e., the marginal cost schedule to the firm buying goods from competitive producers with supply schedule S) is constructed. Prices are determined where the MO schedule crosses the sum of the marginal revenue schedules, $MR_E + MR_A$. The firm will pay price P_S^1 to

[7]A marketing firm in reality incurs costs of transportation, etc.; but if these costs are constant at the margin, Samuelson (1952) has shown that conclusions in this kind of model are not substantially altered—one merely has to shift the axis of one country or the other by the amount of the marginal marketing cost. If marketing costs are not constant, then the results obtained herein will be altered; but often the changes will be minor. In many cases marketing costs are insignificant relative to the value of the good traded. For example, the Canadian Wheat Board charges less than 1 percent of the total value for marketing costs.

producers and charge price P_A^1 to consumers in A and price P_E^1 to consumers in B. Note that the consumer and producer prices are not identical in the exporting country.[8]

Even though large marketing firms may have market power, they are often regulated by government price controls imposed in the domestic market. In order to take price controls explicitly into account, the above model is modified. Since, in the Lerner solution and in the optimal tax situation under competition, producer and consumer prices in the domestic market are identical, and since the general welfare-maximization conditions for country A in Section I indicate domestic producer and consumer prices, the constraint of equality of domestic producer and consumer prices is explicitly considered.[9] Using such a policy, one might expect that domestic consumers are not being exploited when the marketing firm is free to choose its pricing policy abroad. But as shown below, even with this type of control, the marketing firm does not choose optimal export quantities which maximize social welfare in country A.

In the model in this section, the activities of the marketing firm are regulated so that the price charged domestic consumers has to be equal to the price paid to producers. Such a price control could thus be simply monitored by comparing observed market prices for producers and consumers. Since both producers and consumers operate competitively, this control in effect equates domestic consumer price with the marketing firm's average (purchasing) cost which is also equal to the producers' marginal cost (specified by the supply curve). Hence, the marketing firm cannot make profit in the domestic market; its only profit is obtained in the foreign market.

[8]One way to reach the optimum for country A outlined in the previous section is to impose two subsidies, for example, one on domestic production and one on domestic consumption (the Appendix gives the specific subsidies)

[9]Other types of constraints not dealt with here could also be considered (for example, a regulation on the rate of return from goods sold in the domestic market); see Harvey Averch and Leland Johnson; William Baumol and Alvin Klevorick.

The solution (in Figure 1) is obtained by deriving the positive excess supply curve (the difference between domestic demand and supply at different prices labeled as ES) and determining where the marginal schedule (MO_E) to this curve crosses the excess marginal revenue curve (MR_E) of the importing country. As shown by Schmitz and Just under linearity and, assuming S_A intersects the vertical rather than the horizontal axis, the marginal outlay curve (MO_E) to the excess supply curve (ES) crosses the marginal revenue schedule of country B at the *same* level as the MO schedule crosses the sum of the marginal revenue schedules. Hence, exports may well be the same whether or not price controls are imposed domestically.

Although the consumer and producer prices in country A are identical with controls, the optimal allocation of resources is not achieved. The price P_A^2 is *below* the optimal domestic price P_A^* and, with the price constraint, the firm exports too little.[10] This happens because the firm with monopoly-monopsony power disregards the increase in domestic welfare (from increasing domestic production and consumption) when it operates according to the marginal outlay curve associated with the excess supply curve (which is used in country A's welfare maximization) rather than the excess supply curve itself.[11]

To achieve optimality, the government in A has to impose another control in addition to equating producer and consumer prices. One such policy would be to introduce an export *subsidy*. The exact subsidy which achieves optimality for country A can be computed mathematically (see the Appendix) or geometrically (see Figure 1). Geometrically, it is the distance ab since, with this subsidy, MO_E^1 (the marginal outlay after the subsidy) crosses the MR_E schedule at the optimal export level under competition.

[10]Interestingly, however, it can be noted that consumers in the exporting country prefer the regulated monopoly-monopsony to both free trade with competition and social optimality from the standpoint of country A.

[11]This proof simply relies on the fact that the marginal excess supply curve always lies above the excess supply curve.

III. Producer Cartels (or Simple Monopoly)

Consider the alternative case where producers have monopolistic power both domestically and abroad and maximize producer profits from the production of good x. As indicated above, common institutions which can lead to this type of market power are producer cartels or producer marketing boards. As indicated in the Appendix (and Figure 2), the optimal solution for the marketing board is to equate S with the horizontal sum of MR_E and marginal revenue schedule MR_A in country A. The marketing board will restrict producer supply in country A to q^3. The board will charge price P_A^3 to consumers in country A and price P_E^3 to consumers in country B.[12] On the other hand, the optimal trade solution (from the perspective of country A) is determined where ES intersects MR_E. Total production under a marketing board is actually smaller than in the competitive optimal tax case because the horizontal sum of MR_E and MR_A is always below the horizontal sum of MR_E and D_A. Furthermore, exports with the marketing board must be greater than is socially optimal for country A because $MR_E + MR_A$ intersects S_A at a lower price than $MR_E + D_A$; hence, more goods are exported at the lower foreign price corresponding to a lower point on the MR_E schedule. Also, note that domestic producer and consumer prices are not generally identical with the marketing board as under social optimality for country A.[13]

Suppose that, as in the monopoly-monopsony case, the government in A decides to protect local consumers by imposing a constraint similar to the one imposed in the previous case, namely, that the price the consumer pays must be equal to the marginal cost of domestic production. That is, the board can do whatever it wants in pricing the good abroad, but it is regulated in its domestic activities. The results in the Appendix clearly show that the regulated cartel/monopoly exports too much to achieve optimality.[14] This

[12]Of course, it is assumed that no controls are imposed by country B and that competition prevails within country B.

[13]In the absence of domestic demand, one can derive this result from Johnson's analysis (1968) that an export marketing board which maximizes returns to producers in the exporting country attains social optimality.

[14]Comparing the firm optimization conditions of the regulated cartel/monopoly with the social optimization conditions, this result is evident since the second right-hand term in the second condition for the former case is positive.

FIGURE 2. FIRM WITH MONOPOLISTIC POWER-MAXIMIZING PRODUCER RETURNS

is because the regulation prevents the board from exploiting domestic consumers; hence, compared to optimality, the board will sell less locally and will concentrate on exploiting foreign markets. Since the board exports too large a quantity, optimality for country *A* is achieved by imposing an export tax along with domestic price controls.

To see that an export tax is necessary in the producer's cartel case, suppose that the cartel is producing and exporting optimal quantities from the social point of view. If its marginal cost is upward sloping, when it produces one more unit, it can increase the price charged on domestic sales by the corresponding increase in its marginal cost and dump the extra unit produced plus the reduction in domestic consumption on the foreign market. This it will do unless it is stopped from exporting via the tax. Thus, the firm will always tend to produce and export too much unless stopped from so doing by an export tariff.

One must consider however, that an export tax may be difficult to impose. For example, the U.S. Constitution (Article I, Section 9, Paragraph 5) prohibits the United States from restricting trade by use of export taxes. A desirable alternative is to abolish price controls for monopolies and producer cartels when export taxes cannot be imposed; in this case social optimality can be attained by simply imposing a subsidy on domestic consumption (the Appendix gives the unregulated cartel/monopoly case). An advantage of this policy is that is may be less likely to lead to retaliation by trading partners since foreign trade policy is concealed in domestic controls. Explicit taxes and subsidies on exports may induce retaliation by other countries which reduce gains from international trade.[15]

[15]For similar reasons, it may also be advisable to conceal export subsidies for the monopoly-monopsony case However, in this case it is somewhat more difficult to establish adequate alternative controls. For example, if one sets both domestic producer and consumer prices equal and at the appropriate absolute level, there is no guarantee or incentive for the monopoly-monopsony to satisfy domestic demand; it can make the same profit by merely buying domestically and selling abroad. Likewise, if marketing quotas are established, the monopoly-monopsony would have no incentive to use all of its domestic quotas, particularly if domestic prices are controlled. It seems that any adequate alternative controls would need to force the monopoly-monopsony to do at least some minimum level of business with domestic concerns (while also regulating domestic prices). The political and practical feasibility of the latter case may be considerably limited.

IV. Conclusions

A large part of international trading activities (i.e., buying the commodities in producing centers and making the goods available at consuming centers) is carried out by private trading firms and marketing boards. At the same time, federal governments are increasingly regulating the amounts of goods these firms can export and the prices they can charge consumers in the market where the export supply originates. This paper shows that, if the activities of international marketing institutions result in noncompetitive pricing, domestic price controls (where the government sets domestic prices equal to the marginal cost of production) do not result in the optimal allocation of resources even though some groups benefit from resource misallocation. Additional controls (export taxes or export subsidies) are needed along with this type of price control, and the type of fiscal policy depends on the type of marketing institution. With marketing boards, an export tax is suggested; and with the monopoly-monopsony case, an export subsidy is needed. Alternatively, with marketing boards, optimality can be achieved by eliminating price controls and using a simple domestic consumer subsidy.

APPENDIX A: SOCIAL OPTIMIZATION

The Appendix develops the results of the paper mathematically. The first section examines conditions for social optimization independent of the marketing institutions. These results are then used to determine optimal controls for the monopoly-monopsony case (Appendices B and C) and, secondly, for the producer cartel case (Appendices D and E). In each case optimal controls are developed for situations where price controls are not used and then where price controls (equating domestic producer and consumer prices) are used.

Consider an exporting country with inverse competitive supply $S(q)$ and inverse competitive demand $D_A(q_A)$ where q is quantity supplied and q_A is quantity consumed domestically. The importing country has inverse competitive demand $D_E(q_E)$ where q_E is the quantity traded ($q = q_A + q_E$). The social optimum for the exporting country is attained by maximizing the sum of domestic producer and consumer welfare,

$$\max_{q_A,q_E} \int_0^{q_A} D_A(x)\,dx - \int_0^q S(x)\,dx + q_E D_E(q_E)$$

The resulting optimality conditions (assuming second-order conditions hold) are

$$\text{(A1)} \qquad D_A(q_A^*) = S(q^*)$$

$$\text{(A2)} \qquad MR_E(q_E^*) = S(q^*)$$

where * denotes optimal quantities and $MR_E(q_E)$ is the marginal revenue associated with the foreign demand for the export good,

$$MR_E(q_E) = D_E(q_E) + q_E D'_E(q_E)$$

If marketing is carried out under competition in equilibrium, the optimality conditions in (A1) and (A2) are attained, respectively, when $P_A^* = P_s^*$, and $P_E^*(1 + 1/\eta_E) = P_s^*$; where P_A^* is domestic consumer price and P_s^* is domestic producer price in the exporting country, P_E^* is consumer price in the importing country, and η_E the elasticity of demand in the importing country is defined as follows:

$$\eta_E = \frac{P_E}{q_E D'_E(q_E)}$$

Thus, the optimal export tax which attains optimality under competition is $-P_E^*/\eta_E$.

Appendix B: Unregulated Monopoly-Monopsony

Consider now the case of a profit-maximizing marketing firm which possesses both monopoly and monopsony power in both the exporting and importing countries; its objective is to

$$\text{(A3)} \quad \max_{q_A,q_E} q_A D_A(q_A) + q_E D_E(q_E) - q\,S(q)$$

First-order conditions imply

$$\text{(A4)} \qquad MR_A(q_A^1) = MO(q^1)$$

$$\text{(A5)} \qquad MR_E(q_E^1) = MO(q^1)$$

where the superscript 1 denotes optimal quantities for the unregulated monopoly-monopsony case, $MR_A(q_A)$ represents marginal revenue from domestic demand

$$MR_A(q_A) = D_A(q_A) + q_A D'_A(q_A)$$

and $MO(q)$ represents the marketing firm's marginal outlay

$$MO(q) = S(q) + q\,S'(q)$$

The conditions in (A4) and (A5) can be translated into respective price conditions

$$P_A^1(1 + 1/\eta_A) = P_s^1(1 + 1/\eta_s)$$

$$P_E^1(1 + 1/\eta_E) = P_s^1(1 + 1/\eta_s)$$

where η_A and η_s are the elasticities of domestic demand and supply, respectively:

$$\eta_A = \frac{P_A}{q_A D'_A(q_A)} \qquad \eta_s = \frac{P_s}{qS'(q)}$$

To attain social optimality in this case, a production subsidy of p_s^1/η_s and a domestic consumption subsidy of $-P_A^1/\eta_A$ (note $\eta_A < 0$) can be established. In this case, the conditions in (A4) and (A5) become

$$\text{(A6)} \quad MR_A(q_A^1) - \frac{P_A^1}{\eta_A} = MO(q^1) - \frac{P_s^1}{\eta_s}$$

$$\text{(A7)} \quad MR_E(q_E^1) = MO(q^1) - \frac{P_s^1}{\eta_s}$$

since the domestic demand and supply curves are shifted by the amounts of the subsidies. Translating (A6) and (A7) into price terms, however, obtains price conditions $P_A^1 = P_S^1$, and $P_E^1(1 + 1/\eta_E) = P_s^1$, which corresponds to the social optimization conditions in Appendix A.

Appendix C: Regulated Monopoly-Monopsony

Turn now to the case where a monopoly-monopsony marketing firms is regulated by a price control which forces him to equate

domestic producer and consumer prices (in the exporting country). The firm's profit-maximization problem is thus the same as in (A3) except that the appropriate constraint is added

$$\max_{q_A, q_E} q_A D_A(q_A) + q_E D_E(q_E) - q\,S(q)$$

subject to $D_A(q_A) = S(q)$

Using a Lagrangian multipler approach, one obtains necessary conditions

$$\text{(A8)}\qquad D_A(q_A^2) = S(q^2)$$

$$\text{(A9)}\qquad MR_E(q_E^2) = MO(q^2) - [q_A^2 D'_A(q_A^2) - q^2 S'(q^2)]S'(q^2) \div [D'_A(q_A^2) - S'(q^2)]$$

where the superscript 2 denotes optimal quantities for the regulated monopoly-monopsony case. Equations (A8) and (A9) translate into respective price conditions $P_A^2 = P_s^2$, and

$$P_E^2\left(1 + \frac{1}{\eta_E}\right) = P_s^2\left(1 + \frac{q_E^2}{\eta_s q^2 - \eta_A q_A^2}\right)$$

Following the same approach as relating to equation (A6) and (A7), one can verify that imposition for an export subsidy of $P_s^2 q_E^2/(\eta_s q^2 - \eta_A q_A^2)$ attains social optimality for the exporting country in this case.

Appendix D: The Unregulated Producer Cartel

This section considers the case where the marketing institution selling to competitive consumers in both countries is a producer cartel acting on behalf of producers to maximize their just welfare (joint profits),

$$\text{(A10)}\qquad \max_{q_A, q_E} q_A D_A(q_A) + q_E D_E(q_E) - \int_0^q S(x)\,dx$$

First-order conditions imply

$$\text{(A11)}\qquad MR_A(q_A^3) = S(q^3)$$

$$\text{(A12)}\qquad MR_E(q_E^3) = S(q^3)$$

where the superscript 3 denotes optimal quantities in the unregulated producer cartel case. The conditions in (A11) and (A12) translate into respective price relationships

$$P_A^3(1 + 1/\eta_A) = P_s^3; \quad P_E^3(1 + 1/\eta_E) = P_s^3$$

Following the approach in the previous two sections, one can verify that imposition of a domestic consumption subsidy of $-P_A^3/\eta_A$ will cause the cartel to reach social optimality in this case.

Appendix E: The Regulated Producer Cartel

Finally, consider operation of a producer cartel under the regime of price controls forcing equality of domestic producer and consumer prices (in the exporting country). The maximization problem for the cartel is the same as in (A10) with the exception that the constraint must be added.

$$\max_{q_A, q_E} q_A D_A(q_A) + q_E D_E(q_E) - \int_0^q S(x)\,dx$$

subject to $D_A(q_A) = S(q)$

First-order conditions from the Lagrange constrained-maximization problem imply

$$D_A(q_A^4) = S(q^4)$$

$$MR_E(q_E^4) = S(q^4) - \frac{q_A^4 D'_A(q_A^4) S'(q^4)}{D'_A(q_A^4) - S'_A(q^4)}$$

which can be translated into respective price conditions $P_A^4 = P_s^4$, and

$$P_E^4\left(1 + \frac{1}{\eta_E}\right) = P_s^4\left(1 - \frac{q_A^4}{\eta_s q^4 - \eta_A q_A^4}\right)$$

where the superscript 4 denotes optimal quantities in the regulated producer cartel problem. Again, following the approach of the previous sections, one finds that an export tax of $q_A^4 P_s^4/(\eta_s q^4 - \eta_A q_A^4)$ will induce the regulated producer cartel to attain social optimality.

REFERENCES

H. Averch and L. L. Johnson, "Behavior of the Firm under Regulatory Constraint," *Amer.*

Econ. Rev., Dec. 1962, *52*, 1053–69.

W. J. Baumol and A. K. Klevorick, "Input Choices and Rate-of-Return Regulation: An Overview of the Discussion," *Bell J. Econ.*, Autumn 1970, *1*, 162–90.

T. J. Bertrand, "Optimal Tariff Policy Designed for Governmental Gain," *Can. J. Econ.*, May 1973, *6*, 257–60.

W. M. Corden, *Trade Policy and Economic Welfare*, New York 1974.

J. M. Currie, J. A. Murphy, and A. Schmitz, "The Concept of Economic Surplus and Its Use in Economic Analysis," *Econ. J.*, Dec. 1971, *81*, 741–99.

Vernon C. Fowke, *The National Policy and the Wheat Economy*, Toronto 1957.

Jim Hightower, *Eat Your Heart Out: Food Profiteering in America*, New York 1975.

H. G. Johnson, "Optimum Welfare and Maximum Revenue Tariffs," *Rev. Econ. Stud.*, Oct. 1951, *19*, 28–35.

———, "Alternative Maximization Policies for Developing Country Exports of Primary Products," *J. Polit. Econ.*, May/June 1968, *76*, 489–93.

A. P. Lerner, "The Concept of Monopoly and the Measurement of Monopoly Power," *Rev. Econ. Stud.*, June 1934, *1*, 157–75.

P. A. Samuelson, "Spatial Equilibrium and Linear Programming," *Amer. Econ. Rev.*, June 1952, *42*, 283–303.

———, "An Exact Hume-Ricardo-Marshall Model of International Trade," *J. Int. Econ.*, Feb. 1971, *1*, 1–18.

A. Schmitz and R. E. Just, "Semiprice Discrimination," *Econ. Rec.*, Dec. 1977, *53*, 559–74.

E. Tower, "Ranking the Optimum Tariff and the Maximum Revenue Tariff," *J. Int. Econ.*, Feb. 1977, *7*, 73–80.

Factor-Market Distortions and Dynamic Optimal Intervention: Comment

By EDWARD J. RAY*

Recently in this *Review*, Harvey Lapan developed a dynamic analysis of distortions in domestic labor markets. His primary contribution was to point out that the static solution to the problem of distortions is incompatible with optimal adjustment to long-run equilibrium. Lapan concluded that labor market distortions could be handled optimally by providing employment subsidies to firms in the depressed sector that are somewhat less than the employment subsidies implied by static analysis. Unemployment would exist and serve as a policy instrument to encourage labor migration from the depressed sector to the rest of the economy.

In contrast, I will argue that optimal intervention should consist of two elements: a subsidy to employment in the depressed area exactly equal to the static optimal subsidy; and transfer payments to workers to cover the costs of migration from the declining sector of the economy. With this program there would be no unemployment in the depressed area in the short run. The fundamental point I wish to make is that in a dynamic setting, optimal intervention requires the use of two policy instruments, not one. The optimal solution entails both an offset to existing short-run distortions, and a replication of the dynamic path the economy would follow if markets were perfect.

I. The Model

I begin with a standard one-factor, two-sector model.[1] Let M and A be the two commodities, and let N_m and N_a represent labor used in the production of M and A, respectively. Production functions are represented as follows:[2]

$$(1) \qquad M = F_m(N_m)$$

$$(2) \qquad A = F_a(N_a)$$

and $\qquad F_i' > 0, F_i'' < 0 \qquad (i = a, m)$

where F' and F'' refer to first and second partials with respect to labor inputs.

Let A be the numeraire, and let P represent the relative price of M. Furthermore, assume that the number of potential workers in each sector L_i is fixed at any moment in time, and that the total supply of labor is fixed for the period in question.

$$(3) \qquad N_i \leq L_i \qquad (i = a, m)$$

$$(4) \qquad L_a + L_m = \bar{L}$$

where $\bar{L}$ is the fixed aggregate supply of labor. Except for institutional constraints to be discussed shortly, product and factor markets are assumed to be perfectly competitive.

Let us assume that factor and product markets do not adjust instantaneously to exogenous changes in prices and/or employment opportunities, because there exist nonzero adjustment costs that are an increasing function of the speed of adjustment itself. In particular if sector A is expanding and sector M is declining, we have the following adjustment process in the labor markets:

$$(5) \qquad DL_a = \phi(u_m, w_a - w_m)L_m$$

where DL_a is the time rate of expansion of employment in A at time t, u_m is the unemployment rate in sector M at time t, and $\bar{w} = w_a - w_m$ is the wage differential if any at time t. The ϕ function is assumed to have the

*Professor of economics, Ohio State University. I am grateful to Donald O. Parsons and to anonymous referee for numerous helpful comments.

[1]Except for explicit introduction of the labor transfer function and costs of transfer, the model presented here is identical to the model used by both Lapan and John Harris and Michael Todaro. The model is equivalent to the two-factor case with capital immobility.

[2]Detailed derivations are available on request from the author.

following properties:

$$\phi_1 > 0;\ \phi_{11} < 0;\ \phi_2 > 0;\ \phi_{22} > 0;\ \phi(0,0) = 0$$

where ϕ_1 and ϕ_{11} are the first and second partials of ϕ with respect to u_m, and ϕ_2 and ϕ_{22} are the first and second partials of ϕ with respect to $\overline{w}$. The assumption that $\phi(0,0) = 0$ implies that if labor is fully employed and wages are equalized across the economy there is no incentive for workers to move, i.e., $DL_a = 0$.

In his discussion of the dynamic labor adjustment process Lapan assumed that the government was forced by political pressures, etc. to equate wages across sectors of the economy at each point in time. The combination of an arbitrary wage rule and nonzero labor transfer costs is sufficient to generate short-run unemployment in response to exogenous commodity price changes. Given the institutional constraint on wages, Lapan derived a constrained optimal control solution for the adjustment to long-run equilibrium along the lines suggested below.

Assume that we are dealing with a "small," open economy which exports A and imports M and that there is a fall in P, associated with an improvement in the country's terms of trade, to P^*. Our labor market assumptions and initial conditions imply the following expressions:

$$(6)\quad N_a = L_a,\ N_m = L_m(1 - u_m),\ u_m \in [0,1]$$

$$(7)\qquad F_a'(L_a(0)) > P^*F_m'(L_m(0))$$

where $L_a(0)$ and $L_m(0)$ are initial labor supplies in sector A and sector M.

Let T represent the point in time when relocation of labor is completed and full-employment GNP is generated under the guidance of an optimal intervention scheme. Clearly, the value of the constrained path for the economy cannot exceed the value of the unconstrained path along which wages can vary across sectors. The objective of the policy planner will be to satisfy the institutional constraint at minimum economic cost, or to maximize V, where:

$$(9)\quad V = \int_0^T [F_a(L_a) + P^*F_m(N_m) - C(DL_a)]\, e^{-rt}dt$$

where L_a, N_m, and DL_a are the policy induced levels of employment in sectors A and M and the induced rate of labor transfer from M to A. The cost of labor transfer function $C(DL_a)$ is assumed to have positive first and second derivatives. Without the increasing cost assumption with respect to the C function, labor transfers would be complete within one period. Increasing marginal transfer costs are assumed to be caused by time-related limits on the speed with which labor can be absorbed into the A sector. In effect, the more rapidly workers attempt to transfer from the M sector to the A sector, the more rapidly the congestion costs of assimilating them into A will rise.

Lapan's approach is to solve the optimal control problem in one state variable $L_a(t)$ and one control variable u_m. The Hamiltonian would be

$$(10)\quad H = [F_a(L_a) + P^*F_m(N_m) - C(DL_a)]e^{-rt} + \lambda(L_m\phi(u_m, 0))$$

where λ is the present discounted value of transferring a worker from M to A at time t (discounted back to time zero). Defining $q(t) = \lambda(t)e^{rt}$ as the value of transferring a worker at time t evaluated at time t we find that along the optimal path:[3]

$$(11)\qquad q(t) = \frac{P^*F_m'}{\phi_1} - C'$$

In static analysis, the optimum subsidy rate $s(t)$ would be just sufficient to sustain full employment in the short run:

$$(12)\quad s(t) = 1 - [P^*F_m'(L_m)/F_a'(L_a)] > 0$$

The dynamic subsidy rate depends upon the unemployment rate:[4]

$$(13)\quad s^*(t) = 1 - [P^*F_m'(N_m)/F_a'(L_a)]$$

Clearly, $s^*(t) < s(t)$.

[3]Except for the explicit inclusion of the transfer cost term this is the condition derived by Lapan.

[4]The derivation of sufficient conditions for the nonnegativity of s^* is available upon request from the author.

II. The Competitive Market Solution

My basic thesis is that the forced equalization of wages across the economy introduces two separate distortions that require the use of two separate instruments to achieve an optimal solution. First, unemployment is generated in the current period, and second, the wage incentive for labor to shift optimally over time from M to A has been destroyed. To the extent that labor responds to differences in expected wages (reflected in different probabilities of employment in the two sectors at the policy imposed wage rate), the optimal subsidy in Section I was that which provided the best tradeoff between current gains associated with getting workers reemployed in sector M this period vs. future gains associated with reemployment of labor in section A in response to the current unemployment stimulus in sector M. Alternatively, I argue that the optimal solution will consist of a subsidy to employment in sector M that is exactly equal to the optimal static subsidy implied by earlier work and associated with short-run full employment, plus a direct subsidy to labor transfers from M to A (perhaps travel vouchers) that is equal at the margin to the implicit difference between the discounted value of the marginal product of labor in A and M and to the marginal cost of labor transfer at each point in time.

The optimality of the solution just proposed can best be illustrated by summarizing the unconstrained competitive market response to a change in the terms of trade. The unconstrained market response would be equivalent to solving an optimal control problem involving one state variable L_a, and one control variable $\bar{w} = w_a - w_m$. The appropriate Hamiltonian is

$$H = [F_a(L_a) + P^*F_m(L_m) - C(DL_a)]e^{-rt} + \lambda(L_m\phi(0, \bar{w})) \tag{14}$$

Optimizing the Hamiltonian with respect to $\bar{w}$ yields a dynamic path of adjustment along which there is no unemployment in the declining sector and that implies the following contemporaneous marginal value for labor transfers from M to A:

$$q(t) = C'(DL_a) \tag{15}$$

The exhaustive study of static domestic distortion problems has taught us that the optimal solution corresponds to the use of that instrument which exactly offsets or neutralizes the impact of the distortion on the economy without introducing any new distortions. Applying that same reasoning to dynamic distortion problems, we should employ that set of instruments which satisfies the institutional constraint that wages be equalized and yet also replicates the short- and long-run behavior of the unconstrained economy.

Specifically, we should equalize wages in the short run by providing a subsidy S, equal to the optimal static solution subsidy which is equal to the difference between the value of the marginal products of labor in the two sectors at full employment, i.e.,

$$S = F'_a(L_a) - P^*F'_m(L_m) \tag{16}$$

The fact that $\phi(0, 0) = 0$ means that without further intervention there will be no transfer of labor from M to A. From equation (15) we know the unconstrained competitive market condition for optimal labor transfers. Therefore, a subsidy, $S_{\bar{w}}$ for labor transfers from M to A that equates q with the marginal cost of labor transfer in each period would replicate the optimal rate of labor reallocation observed in the unconstrained case. In summary, we need two subsidies to promote full employment in the short run and optimal labor transfers in the long run.

REFERENCES

J. Bhagwati and T. N. Srinivasan, "On Reanalyzing the Harris-Todaro Model: Policy Rankings in the Case of Sector-Specific Sticky Wages," *Amer. Econ. Rev.*, June 1974, *64*, 502–08.

J. Harris and M. Todaro, "Migration, Unemployment, and Development: A Two-Sector Analysis," *Amer. Econ. Rev.*, Mar. 1970, *60*, 126–42.

H. E. Lapan, "International Trade, Factor Market Distortions, and the Optimal Dynamic Subsidy," *Amer. Econ. Rev.*, June 1976, *66*, 335–46.

Factor-Market Distortions and Dynamic Optimal Intervention: Reply

By HARVEY E. LAPAN*

Edward Ray, in his comment on my 1976 paper, analyzes a slightly different model than the one I presented, and thus reaches different conclusions. His principal conclusions are that: (i) given wage rigidities, a wage subsidy to producers is needed, and this subsidy is equivalent to the optimal static subsidy that ensures full employment in each sector; and (ii) given the forced equilization of wages across sectors, a subsidy to workers is needed to encourage labor transfers between sectors. Thus, Ray finds that full employment is always desirable, whereas I find that some unemployment is (usually) present along the optimum path.

The differences in our solutions arise from the different specifications of our models. Ray assumes (his notation):

(1) $DL_a = \phi(u_m, w_a - w_m)L_m$

(2) $C = C(DL_a) =$ cost of labor transfer

Thus, from (1), Ray assumes that the rate of labor transfers between sectors depends on the unemployment rate in the declining sector (u_m), and the differential in wages received by workers between the two sectors ($w_a - w_m$). Consequently, (1) represents a behavioral relation concerning labor's voluntary decision to move between sectors. Furthermore, Ray assumes that labor transfers are possible even under full employment; i.e., $\phi(0, w_a - w_m) > 0$ for $w_a > w_m$. Since the return to labor is treated as a pure rent (i.e., the labor supply decision is not responsive to the wage rate) and since Ray assumes that any desired labor transfer between sectors can be accomplished without unemployment (by properly choosing $(w_a - w_m)$), it immediately follows that full employment is always desirable.[1]

In order to prevent the optimal long-run solution from being obtained costlessly and immediately, Ray assumes that the process of transferring labor is costly $[C = C(DL_a)]$. Though this cost function is not explained by Ray, the costs obviously do not represent the value of output foregone due to (voluntary or involuntary) unemployment; rather, these costs depend only on the number of workers transferred, and not on how the transfer is accomplished.

Thus, the key relationship in Ray's model is this cost function, as can be seen from his equation (9). The optimal rate of labor transfer for a centrally planned economy is determined from (9), and this optimal solution is independent of any distortions that occur within the economy. Of course, how this plan is supported by prices and subsidies will depend upon the types of distortions present, but the conclusion that full employment is always desirable derives solely from his specification of the control model, and not from the range of policy options available to planners who are attempting to achieve this optimal plan.

In my specification of the control problem (1976) it is assumed that some unemployment is necessary in order to transfer labor between sectors:

(3) $DL_a = \phi(u)L_m$

Again, (3) does not imply the presence of distortions, but rather specifies how transfers are accomplished; for example, some search may be necessary before new employment is obtained. Given (3), the optimal rate of unemployment and labor transfer for a centrally planned economy is determined; the policy needed to support it then depends on the types of distortions present. As noted in

*Professor of economics, Iowa State University.

[1] Presumably, the institutional constraint is the forced equalizaiton of wages across sectors: $w_a = w_m$. However, implicitly Ray assumes that those workers who do move between sectors can receive an extra subsidy to compensate them for the extra costs they incur.

the comment by James Cassing and Jack Ochs and in my reply (1978), this unemployment may be voluntary or involuntary; and whether private decisions are socially optimal depends not only on whether price rigidities are present, but also on how private labor transfers are made and on whether congestion occurs in the search process. It is clear that (3) represents the implicit cost of labor transfer; explicitly, the cost of labor transfer is the value of output lost through unemployment:

$$(4)\ C(DL_a) = P_m\,[F_m(L_m) - F_m(L_m(1-u))] = P_m[F_m(L_m) - F_m(L_m g(\frac{DL_a}{L_m}))]$$

$$(5)\quad \phi(u) = (DL_a/L_m); \quad u = \phi^{-1}(DL_a/L_m); \quad (1-u) = g(DL_a/L_m)$$

Thus, the cost of labor transfers that Ray postulates can be derived from the assumption that some unemployment is necessary to accomplish labor transfers. Therefore, whether any unemployment is desirable depends only upon the mechanism by which labor is transferred between sectors.

Finally, how the optimal plan is supported depends on the type of distortions present. As Cassing-Ochs and I (1978) show, if no wage rigidities are present (so that all unemployment is voluntary) and if individuals have perfect foresight, then private decisions will be socially optimal if, and only if, no congestion occurs in the search process; if congestion occurs, then some intervention is needed to support the optimal plan. However, if wage rigidities are present then, as I argued in my earlier paper, a wage subsidy to producers is required in order to provide for the optimal rate of employment (unemployment) in the declining sector. Furthermore, as long as some unemployment is required to transfer labor, this subsidy will be less than the optimal static subsidy, as described in my earlier paper. Ray's conclusion that the optimal dynamic subsidy to producers is the same as the optimal static subsidy does not hold if unemployment is required to affect transfers.

Furthermore, if the only private costs of search to workers are the wages foregone, then all unemployed workers in the declining sector (as a result of the dynamic subsidy being less than the optimal static subsidy) will find it profitable to search for work in the other sector, and no additional subsidy to workers will be required to support the optimal plan.[2] Thus, if at the beginning of each day, the government announces the subsidy for that day, and then the firms tell which workers to show up for work, those workers who find themselves unemployed will choose to search for a job in the other sector.[3] Hence, only the subsidy to producers is required to support the optimal plan.

To conclude, Ray finds that full employment is always desirable, and that two subsidies—one to producers and one to workers—are required. His first conclusion differs from my conclusion because he uses a different model; the specification of the control model, and not the presence of distortions, determines whether full employment is optimal. Regarding his second conclusion, we have seen that the policies needed to support the optimal plan depend on the types of distortions present and the way in which the labor market functions. For his model, since full employment is desirable, his conclusions are correct; however, for my specification, only a subsidy to producers is needed since the only distortion present is the wage rigidity. Finally,

[2]This is true since it is assumed that wages received in the two sectors are identical. Assume that the worker cannot be employed in the more productive sector on the first day; if he searches for the day, he has probability $\phi(u)$ of finding a job there, and hence being employed for subsequent days. If there is some positive probability he will not get his old job back, then as long as $\phi(u) > 0$ search is desirable for him. Of course, if the search process itself entails costs—above the wages foregone—then he may not choose to search. But if this search process does entail these costs, they should be included in the original control problem.

[3]Of course, if the workers can only visit one firm per day, and if they do not know *ex ante* whether their old job is available for that day, they must decide whether to show up for work, hoping to be employed, or whether to look for a job elsewhere. In this case, the optimal subsidy derived in my original paper will not be sufficient. However, note that this specification is not consistent with the control model since the latter implictly assumes search is needed only to find jobs in the other sector. If search is needed in both sectors, the control model must be modified (see my 1978 paper, fn. 4).

note that both models allow the attainment of the first best solution (given the *assumption* that labor mobility is costly), despite the presence of distortions. If labor supply decisions were endogenous, and if wages received were required to be the same in each sector, then the return to labor would not be a pure rent and we would have a true second best world.

REFERENCES

J. Cassing and J. Ochs, "International Trade, Factor-Market Distortions, and the Optimal Dynamic Subsidy: Comment," *Amer. Econ. Rev.*, Dec. 1978, *68*, 950–55.

H. E. Lapan, "International Trade, Factor Market Distortions, and the Optimal Dynamic Subsidy," *Amer. Econ. Rev.*, June 1976, *66*, 335–46.

——, "International Trade, Factor-Market Distortions, and the Optimal Dynamic Subsidy: Reply," *Amer. Econ. Rev.*, Dec. 1978, *68*, 956–59.

E. J. Ray, "Factor-Market Distortions and Dynamic Optimal Intervention: Comment," *Amer. Econ. Rev.*, Sept. 1979, *69*, 715–17.

The Disequilibrium Model in a Controlled Economy: Comment

By BARBARA GOODY KATZ*

In a recent paper in this *Review*, David Howard concludes that his test of the predictive ability of the Barro-Grossman disequilibrium model is successful within the Soviet context. The Barro-Grossman (B-G) model (1971, 1974) focuses on the responses of saving and labor supply to conditions of excess demand, defined as the case where at the prevailing price, the demand for consumer goods exceeds their supply. The B-G model predicts that under conditions of excess demand, referred to also as repressed inflation, an increase (decrease) in the quantity of goods available leads to a decrease (increase) in saving and an increase (decrease) in labor supply, with its consequent multiplier effect on output. Howard tailors the B-G model to the Soviet economy by including an uncontrolled market, specifically the collective farm market, and by eliminating the role of profits as an argument in the effective saving and labor supply functions.

While the B-G model supplies an important theoretical framework for predicting responses to changes in the constrained availability of goods, the model is difficult to test empirically. Moreover the methodology used by Howard is flawed in several ways. An adequate measure of the supply of goods available on the constrained market (defined by Howard as B^0, goods available on controlled market) is needed to test the B-G model. However, Howard's choice for B^0, a composite of goods sold on the state and cooperative retail markets, is not such a measure. The significance of introducing a variable to measure availabilities is to capture the notion that what is purchased under conditions of excess demand will not reflect desired purchases, but rather actual purchases, thereby revealing a point of market disequilibrium. Despite the argument that sales seem a good proxy for availabilities because quantities are so limited that whatever is available will be purchased, the use of sales to represent availabilities is unjustified for two reasons.

The first reason focuses on the relationship between two of the most important variables in Howard's analysis: B^0, state and cooperative retail sales, and s^d, the change in savings bank deposits, which can simply be termed saving in the Soviet context where the purchase of interest-earning assets (or even lottery bonds) and the other usual alternatives to savings bank deposits are negligible. Certainly the change in state and cooperative retail sales (ΔB^0) and the change in saving (Δs^d) must be highly negatively correlated. This can be easily seen in the context of the Howard model by constructing the ratios $\Delta B^0/\Delta(wL^s)$ and $\Delta s^d/\Delta(wL^s)$ where w is the wage rate, L^s is the average nonprivate civilian employment, and the product wL^s is the wage bill. Assuming the wage bill is approximately equal to income, these ratios are approximately the marginal propensity to consume and the marginal propensity to save, which must sum to unity, neglecting the fraction of the change in total income spent on the uncontrolled market and assuming other forms of saving to be zero. Noting this obviously negative relationship between Δs^d and ΔB^0, the fact that Howard's estimations yield coefficients for $\partial s^d/\partial B^0$ that he claims are both correct and significant with respect to one-tailed t-tests (see his Table 2: $\partial s^d/\partial B^0 = -.424641$ with an absolute t-statistic of 2.292 and his Table 3: $\partial s^d/\partial B^0 = -.42054$ with an absolute t-statistic of 2.158) is uninformative except when viewed from a somewhat different perspective; that is, the fact that the coefficients are only so weakly signif-

*Assistant professor of economics, Graduate School of Business Administration, New York University. This work was funded in part by a New York University faculty research grant.

icant indicates difficulties with the model itself. Of further econometric consideration is that the technique used by Howard to reach a solution to his model, three-stage least squares, only has asymptotic properties and not for sample sizes as small as thirteen.

But even if there were no problem involving the relationship between s^d and B^0 there is a second reason why sales are an inappropriate proxy for availabilities. There is a systematic bias incorporated into the sales variable when total output is composed of changing proportions of both highly desired and inferior goods. Over the time period under consideration, Howard notes "conditions were getting better" (p. 871), presumably meaning that the availabilities of at least some highly desired goods on the controlled market were increasing relative to the total quantity of goods available on that market. With more desired goods in the output mix, spending on goods of inferior quality, design, and assortment will decline. This will lead to inventory accumulation despite the presence of generalized excess demand. Spending on highly desired goods will tend to increase, although in many cases with a lag reflecting the necessary postponement of current consumption in an attempt to accumulate funds sufficient for the purchase of the highly desired items.[1] This was precisely the Soviet experience in the early 1960's. Since the existence of unsold stocks is well-known, the use of sales as a proxy for availabilities is not accurate, becoming systematically less reliable as the output mix is regularly altered to favor the more highly desired goods.

Another problem with Howard's empirical test of the B-G model relates to the untested but assumed homogeneity of the period 1955–67. During that time there were, in the main, two regimes, one headed by Khruschev from 1956 until 1964, and the other headed by Brezhnev from 1965 onward. The policies followed by the two governmental leaders presumably were different in many areas, including the provision of goods to the controlled market. Since Howard's test of the B-G model suggests that the most important reaction to repressed inflation occurs in household savings deposits, the focus will be on explaining that observed phenomenon. One possible way of testing the presumption that the savings data belong to two different periods, depending on the governmental regime, is to include a dummy variable with a value of zero in the Khruschev years and a value of one in the Brezhnev years. This is added to the other independent variables that Howard employs in equation (3″), the actually estimated version of his equation (3). Following Howard, this equation is defined as

$$SD = \alpha + \pi_1 B^0 + \pi_2(H_{-1}/P) + \pi_3(w/P) + \pi_4(P^A/P) + \pi_5 D \tag{1}$$

where SD = the change in savings bank deposits divided by P and by working age population

B^0 = estimated real sales on the controlled market divided by working age population

w = nominal wage

P = estimated general price index

H_{-1} = nominal amount of savings carried over from the previous period divided by working age population

P^A = estimated price index of goods on free market where $P = 91.3P^B + 8.7P^A$ and P^B is actual state retail price index

D = dummy, 1956–64 = 0, 1965–67 = 1

Unfortunately, however, in this case simply adding the dummy variable to the list of right-hand variables and estimating the three-stage least squares solution to the revised model is not in itself particularly helpful. The high level of multicollinearity renders the t-statistics imprecise.[2] Not surprisingly, under

[1] The length of the lag would seem to depend on the degree to which the mix alteration was anticipated.

[2] Despite the fact that three-stage least squares is known only to have properties asymptotically, and the sample size here is small, it was necessary to use the same technique as Howard in order to make the results comparable. Since this equation is part of a simultaneous model ordinary least squares would provide inconsistent estimates of the coefficients.

TABLE 1—SIMPLE CORRELATION COEFFICIENTS

	D
SD	.934876
B^0	.775224
P^A/P	−.073908
w/P	.824366
H_{-1}/P	.806561

tnese conditions, the coefficient of the dummy variable, among others, is insignificant. For some insight into the multicollinearity problem see the simple correlation coefficients between the dummy variable and the other right-hand variables in Table 1. The inclusion of the dummy variable also causes the value of R^2 in the equation to be negative. The coefficients and their t-statistics (in parentheses) for the equation including the dummy variable based on the maximum values of P^B, P, and B^0 are:

B^0:	849406.	(.812057)
H_{-1}/P:	−32.0251	(−1.90133)
w/P:	.011557	(2.43701)
P^A/P:	−3.76085	(−2.38662)
D:	−.409424	(−1.92094)
α:	−2.35122	(−1.95859)

The R^2 value is −.8463 and the Durbin-Watson ($D.W.$) d-statistic is 1.7724. Three-stage least squares does not maximize R^2 nor does the procedure bound R^2 between 0 and 1. A negative R^2 is possible when the sum of squared residuals is greater than the total sum of squares. This may happen in two-stage least squares since residuals are defined with respect to y but coefficients are estimated with respect to an estimated value for y, i.e., y^*. In addition, this may happen in three-stage least squares as there is no reason for the sum of squared residuals of a given equation in a model to be less than or equal to the total sum of squares for that particular equation even though for the stacked model in its entirety the sum of squared residuals cannot exceed the total sum of squares. Three-stage least squares estimates of that equation without the dummy variable for the period 1955–67 yield an R^2 of .7713, a $D.W.$ d-statistic of 2.0936 and a t-value for the coefficient of B^0 of −1.60018, somewhat different results than those reported by Howard: R^2 = .898, $D.W.$ = 2.249 and an absolute t-value = 2.158 for the B^0 coefficient (see his Table 3).[3]

Since the high level of multicollinearity in (1) described above renders the t-values imprecise, the presumption that the savings data belong to two different periods is pursued further. Equation (1) above is reformulated in the following way:

$$SD = \alpha + \pi_4(P^A/P) + \pi_5 D \tag{2}$$

and the entire model is reestimated using three-stage least squares with these results for equation (2):

α =	−.022375	(−.365420)
π_4 =	.124135	(1.84221)
π_5 =	.179509	(20.5431)
R^2 =	.8697,	$D.W.$ = 2.8994

The coefficient of the dummy variable is significant in explaining the behavior of SD, leading to the hypothesis that the differing policies of different governmental regimes are mainly responsible for determining the demand for real saving. No formal hypothesis testing of alternative equation specifications, that is, of equations (1) and (2) and Howard's original equation (3″), can be done on single equations embedded in simultaneous models. However, the results of one-on-one OLS regressions of each of the following variables: B^0, w/P, H_{-1}/P, P^A/P, and SD on the dummy variable plus a constant term are impressive and helpful in this connection (see Table 2). These one-on-one OLS regressions, in addition to testing the homogeneity of means, also provide R^2 statistics. Since, as noted, hypothesis tests on individual equations within simultaneous equation models

[3]The results of my reestimation of equation (3″) of Howard's model (without the dummy) for the years 1955–67 show:

B^0:	−364740.	(−1.60018)
H_{-1}/P:	−.700281	(−1.93607)
w/P:	4.12104	(5.23200)
P^A/P:	−1.08220	(−6.01706)
α:	−.545572	(−1.73679)
R^2 = .7713,	$D.W.$ = 2.0936.	

TABLE 2—*OLS* REGRESSIONS OF THE VARIABLES IN EQUATION (3″) ON THE DUMMY AND A CONSTANT

Dependent Variable	Dummy	Constant	R^2
B^0	.00000048	.000001201	.6010
	(3.88084)	(20.8000)	
w/P	1.85181	6.28444	.6796
	(4.60532)	(31.2579)	
H_{-1}/P	.615022	.761585	.6505
	(4.31459)	(10.6855)	
P^A/P	−.009503	.904648	.0055
	(−.234357)	(44.6178)	
SD	.178346	.089908	.8740
	(8.32833)	(8.39701)	

are inappropriate, the knowledge of these R^2 statistics is valuable in indicating the importance of the omission of a variable to represent differing political climates in an equation attempting to explain SD. It is apparent that the dummy variable alone can explain 87 percent of the variance in SD as well as considerable variation in three of the four explanatory variables employed by Howard. That is, the existence of different governmental regimes, for which the dummy variable is a proxy, is able to explain much of the behavior of several economic series: real sales of goods on the state and cooperative markets, real wages and real savings balances, lagged one period. Thus, the data suggest that the exclusion of any variable to represent the different governmental regimes results in an artificial homogeneity during this period that simply was not present.

This analysis also challenges Howard's claim that the most important consequence of repressed inflation is its impact on saving; it would appear that the existence of repressed inflation has very little to say about the demand for real saving, whereas knowing the governmental regimes, and when they changed, is an important explanatory variable for the demand for real saving. Of course, shifts in political power imply, among other things, different attitudes towards the provision of goods to the controlled market.

Another way of interpreting the dummy variable is to see it as a proxy for consumer expectations regarding goods' availability; that is, expectations are assumed to be a function of observed governmental actions which are taken as constant throughout a regime and different between regimes. If expectations increase, that is, if consumers believe more of the highly desired goods will be available, saving will increase ($\partial SD/\partial D$ = .179509 in equation (2) with a t-statistic of 20.5431). A different conclusion follows from the B-G model as tested by Howard: an increase in availabilities is associated with a fall in savings as $\partial SD/\partial B^0 = -.42054$ in his equation (3″) with an absolute t-statistic of 2.158. Studying the saving data used by Howard reveals a roughly eight-fold increase in real savings in the 1955–67 period. In that same time the average monthly real wage bill slightly more than doubled, and the state and cooperative retail sales index increased on the order of two to two and one-half times. That is, savings grew dramatically more than purchases. Continuing the data analysis through 1975 it appears that 1) per capita savings deposits have been growing at about 15 percent per year since the mid-1950's while disposable money incomes and per capita outlays on goods and services have been growing at less than half that rate, and 2) total savings deposits exceeded the target of the Ninth Five-Year Plan by 11 billion rubles in 1975, which is an amount larger than the total 1974 retail sales of nonfood items (see Gertrude Schroeder and Barbara Severin). It is this pronounced growth in savings that has been of concern to Soviet planners for several reasons. Not only do the unplanned and sizeable accumulations suggest serious inadequacies in the planning mechanism (i.e., an obvious inference from the existence of unsaleable stocks and unplanned savings) but they also present a clear destabilizing threat to the economy, and therefore the government, due both to their size relative to the availability of highly desired goods and to their seepage into the "second," or unofficial, economy.

Assuming that availabilities of goods on the state and cooperative markets, although still constrained, were more plentiful in the more recent period, the B-G model as used by Howard would predict a decline in savings;

yet, to the concern of the Brezhnev regime, savings have been increasing. One way to reconcile these observations with the predictions of the model is to recast the model in multiperiod terms where consumers maximize intertemporal expected utility. Consider the following: Consumers know the availability of goods on the controlled market today with certainty and have expectations in the form of subjective probability distributions over differing levels of availabilities of goods tomorrow. Introducing expectations about future availabilities into the disequilibrium model as done here with the inclusion of a dummy variable alters the predictions of the model; expectations about rising future availabilities suggest increased saving (recall $\partial SD/\partial D > 0$ in equation (2) and significant at the .01 level), particularly in an economy where the purchase price of all items must be paid in advance. For example, in the five-year period 1970–75 the number of cars sold to individuals increased eight times (see Schroeder and Severin); since the cost of the Zhiguli automobile was about 7,500 rubles in 1975 and since cash prepayment was required, it follows that if a customer anticipated the supply increase, and wanted to obtain a car, savings would increase with increases in expected availability. Using a dummy variable as a proxy for governmental regimes and associating different regimes with different expectations concerning availabilities is not a substitute for a fully generalized multiperiod disequilibrium model, but it does suggest a way to reconcile the facts of tremendous savings accumulations with the predictions of a disequilibrium model.

REFERENCES

R. J. Barro and H. I. Grossman, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 83–93.

______ **and** ______, "Suppressed Inflation and the Supply Multiplier," *Rev. Econ. Stud.*, Jan. 1974, *41*, 97–104.

______ **and** ______, "Suppressed Inflation and the Supply Multiplier," *Rev. Econ. Stud.*, Jan. 1974, *41*, 97–104.

D. H. Howard, "The Disequilibrium Model in a Controlled Economy: An Empirical Test of the Barro-Grossman Model," *Amer. Econ. Rev.*, Dec. 1976, *66*, 871–79.

G. E. Schroeder and B. S. Severin, "Soviet Consumption and Income Policies in Perspective," Joint Economic Committee, U. S. Congress, *Soviet Economy in a New Perspective*, Washington, Oct. 1976.

The Disequilibrium Model in a Controlled Economy: Comment

By MACHIKO K. NISSANKE*

In his recent work in this *Review* on the application of the Barro-Grossman model to centrally planned economies (*CPE*s), David Howard presents a relatively simple model of disequilibrium in a controlled economy together with some empirical results, which, he claims, strongly support the presence of repressed inflation in the *USSR* (1955–67). On the basis of his results, he confirms the predictive ability of the Barro-Grossman model, and he claims to have improved our knowledge of Soviet repressed inflation.

Similar empirical investigations of other *CPE* countries, Howard also suggests, are of great importance in testing the general applicability of his model and his estimation technique. It is with this aim that a test of the Polish economy (1955–75) was undertaken (see my paper), but certain difficulties encountered in the process of estimation revealed several fundamental drawbacks in his methodology and estimation technique.

The main *technical* problems are: 1) the identification of the model, and 2) the specification of the labor demand function and hence the endogeneity of prices and wages. These will be dealt with first. I shall then go on to discuss more basic problems. These are concerned with Howard's lack of any model of planners' behavior, the assumptions of his maintained hypothesis, and the unjustifiable claims he makes for his results.

I

Howard's model has eight equations:

$$(1)\quad L^s = L^s(B^0, H_{-1}/P, W/P, P^B/P, P^A/P)$$

$$(2)\quad A^d = A^d(B^0, H_{-1}/P, W/P, P^B/P, P^A/P)$$

$$(3)\quad s^d = s^d(B^0, H_{-1}/P, W/P, P^B/P, P^A/P)$$

$$(4)\quad A^s = A^s(B^0, H_{-1}/P, W/P, P^B/P, P^A/P, G, R)$$

$$(5)\quad A^d = A^s$$

$$(6)\quad L^d = L^d(W/P, Q^0, K^0)$$

$$(7)\quad L^s = L^d$$

$$(8)\quad P = \lambda_1 P^A + \lambda_2 P^B$$

where L^s = labor supply
A^d = free market demand
s^d = demand for real saving
A^s = free market supply
L^d = labor demand
B^0 = goods available on controlled market
W = nominal wage
P = general price level
H_{-1} = nominal amount of saving carried over from the previous period
P^B = price level of goods on controlled market
P^A = price level of goods on free market
G = dummy variable for government policy towards agricultural sector
R = dummy variable for weather
Q^0 = net material product
K^0 = capital stock

As the model stands, there is a problem with its identification, even by the order

*Department of economics, Birkbeck College, University of London. This research work was supported by the Social Science Research Council, U.K., under grant HR 4985. I thank Richard Portes for his guidance and valuable comments throughout this study and am very grateful to John Muellbauer, Ron Smith, Hugh Davies, David Winter, and Stan Rudcenko for assistance and valuable discussion on the subject. I am solely responsible for the content of this paper.

condition. Since it is non-linear in P, let us write it in terms of new variables: $HNR = H_{-1}/P$, $WR = W/P$, $PAR = P^A/P$, $PBR = P^B/P$. If we assume all the equations are then linear it appears that the free market supply equation (4) in Howard's original model is not identified. However, by using equation (8) and deleting P^B/P from the system, he removes his problem at the cost of not being able to estimate some parameters without a priori information on the ratio λ_1/λ_2. This elimination is also necessary to avoid multicollinearity caused by the linear relationship between PBR and PAR.

Given the way Howard writes the model, WR and PAR, which are endogenous variables in his specification, are not dependent variables in any equation. It might be better to renormalize either L^s or L^d and A^s or A^d to make them dependent. It should be noted that two-stage least squares (*2SLS*) and three-stage least squares (*3SLS*) estimates are not invariant to the choice of normalization, and it is not obvious that his normalization is a plausible one.

Normalizing the L^d equation by WR and A^s equation by PAR, *3SLS* was used on both the Polish and Soviet data; but in both cases, the residual covariance matrix necessary for the third stage was singular because there was insufficient exogenous variation to identify the system, that is, the rank condition for the system is not fulfilled. Only after adding several extra exogenous variables were *3SLS* estimates obtained. Given this result, it is puzzling how Howard obtained his estimates.

It is possible that he used calculated residuals from the second stage of *2SLS*, which also include the reduced-form residuals, to compute the residual covariance matrix for the third stage. There is, however, no econometric justification for such a procedure, which will produce biased estimates of standard errors. Alternatively, he may have just estimated the first three equations by *3SLS*. But, since the first three equations contain identical regressors, *3SLS* and *2SLS* would then be identical.

The second technical problem in Howard's approach is his treatment of variables as endogenous or exogenous. Howard argues that while in the individual household functions all variables on the right-hand side were exogenous since the individual household is a price taker, W, P^A, and P cannot be taken as given *in the aggregate case* because they are simultaneously determined with other variables L^s, A^d, and s^d. For this reason, the simultaneous system was specified. Endogeneity of W is justified by his view that W is a freely adjusting variable depending on the enterprises' need for labor, rather than being controlled by planners in the light of supply conditions on the commodity market. According to his argument, equilibrium $L^s = L^d$ on the labor market is supposed to be achieved by a wage adjustment process, since, in Howard's words, "the authorities, one way or another, create the funds for wage overexpenditures" (p. 874). This is a manifestation of his belief in the existence of a nearly vertical labor demand function, due to the absence of actual control over wage expenditures. He argues that shortages of consumer goods result in a decrease in the labor supply, but because of the inelastic labor demand function, this results in a higher wage rather than a decrease in labor actually transacted. It follows then that large wage adjustments are required to clear the labor market, if L^d is as inelastic as Howard assumes it to be. In addition, Howard includes the inverse production function *and* the real wage level together as arguments determining labor demand, though he gives no theoretical justification for combining an inverse production function $g^{-1}(Q^0, K^0)$ and the real wage into a simple regression equation. After all, if L^d is in fact vertical, then W/P should not appear at all in this equation, since the coefficient on W/P should be nearly zero.

Although planners actually try to use wage differentiation to attract workers into industries or regions which are given high priority in the current plan, this does not mean that the enterprise is free to change wage levels. Indeed if, for example, the planned wage level is determined *ex ante* by the supply capacity of consumer goods, the behavior of enterprises cannot be properly considered in isolation from planners' decisions. It is an oversimplifi-

cation to say that the wage can serve as an instrument for full and costless adjustment to equilibrium between demand and supply on the labor market.

There also exists the practice of wage control over enterprise activity in the form of bank control on the employment plan and the average wage plan, which restricts expenditure on the wage fund. Historically, while weak wage control was one of the characteristics for the earlier period when open inflation prevailed in East European countries, in the later period (essentially after 1955), the leadership of these countries learned from past experience and began to impose an effective control on the wage fund available to each enterprise in order to keep it within the planned limits (see Franklyn Holzman). In aggregate, the determination of the planned wage fund is based on both the analysis of the balance of income and expenditure in the plan period and on the rate of increase of labor productivity. The degree of success of wages policy varies from country to country and will depend on given demographic conditions. It varies as well from time to time according to the strategic emphasis of a given plan. Detailed analysis and discussion of the history and practice of wage determination and wage control can be found in John Farrell and in Alistair McAuley for the Polish and Soviet cases, respectively. In ignoring completely these considerations, Howard misspecifies the behavior of managers of enterprises and planners and thus his labor demand function and equilibrium condition.

II

Turning to the more fundamental problems with Howard's work, it is clear that there is no consideration of planners' behavior, in spite of the fact that planners play an important role in the determination of overall supply and demand conditions on both commodity markets and labor markets. Howard correctly stresses the importance of simultaneity in the Barro-Grossman model of disequilibrium Then it should not be forgotten that disequilibrium is created and amplified by the interaction in both markets, and hence by the decisions of both planners and households, through multiplier processes of further divergence of the quantity actually transacted from notional demand or supply. This incompleteness is acknowledged by Howard himself, when he admits the necessity of modelling the behavior of policymakers for more detailed analysis (p. 874). If this is so, it is inappropriate to construct a *simultaneous* model representing incorrectly the behavior of the agents on one side of the economy.

Moreover, incorporating an equilibrium condition and estimating a simultaneous model as a whole under this condition is not in the spirit of the Barro-Grossman disequilibrium model. Using Howard's model, one cannot analyze the feedback multiplier effect from the labor market to the consumer goods market without introducing an additional assumption. In view of these shortcomings in Howard's model, it is preferable either to estimate the first three equations by a single equation estimation technique and regard W and P as exogenous variables, or to include planners' reaction functions, if it is desired to allow for simultaneity. The simultaneity arises primarily through planners' responses, not market responses as Howard's model suggests.

Finally, Howard's basic methodology is inappropriate for testing the model. In his methodology, first, he treats the existence of disequilibrium (excess demand) on the commodity market as a part of the maintained hypothesis and then examines three major effects on the labor market, the free market, and on savings behavior. He then includes two equilibrium conditions for the labor market and the free market and uses a simultaneous equation estimation technique, on account of the endogeneity of the aggregate wage and price variables. In other words, rather than *testing* for disequilibrium on each market, he imposes a priori the assumption that there is *disequilibrium* (excess demand) in the commodity market and *equilibrium* in the labor market and then examines the *interaction* between two markets. But for a more elaborate analysis of disequilibrium, the *existence* of disequilibrium on the commodity market itself should be investigated properly. It is the

latter with which we are most concerned, and without careful examination of this aspect, a proper interpretation of results is not possible. Howard's test of the Barro-Grossman model consists of arguing that the signs of the coefficients he estimates support the model. However, since his specification entails incorporating the Barro-Grossman model as a part of the maintained hypothesis, this in no sense constitutes a test.

In fact, as Barbara Katz and Richard Portes and David Winter (1978b) discuss, the positive coefficient of sales on the controlled market (B^0) in his labor supply equation and the negative one of the savings equation are not indicators of the presence of repressed inflation, but natural implications of his definition of the variable B^0. Since total household consumption is made up largely of B^0, the coefficient signs which Howard takes to confirm the hypothesis of excess demand are the same as would be expected if there were no excess demand.

Before reaching a final conclusion about the existence and the extent, if it exists, of repressed inflation in these countries, we have to use disequilibrium estimation techniques, in which the minimum condition ($Q_t = \min(D_t, S_t)$) replaces an equilibrium condition. This is the correct approach to empirical testing of the applicability of the Barro-Grossman model of quantity rationing. Having initially obtained (1977, 1978a) reasonable consumption goods supply and demand functions separately, Portes and Winter (1978b) have applied the methods developed by G. S. Maddala and Forrest Nelson to the consumption goods markets in four *CPE*s (Czechoslovakia, the German Democratic Republic, Hungary, and Poland). This and other studies using disequilibrium methods have so far dealt with only one market rather than interacting markets in disequilibrium, but as pointed out above, Howard also deals essentially with only one market which he *assumes* to be in excess demand.

III

For comparison with Howard's results on the Soviet Union, I present below the results on the Polish economy (1955–75). The first three equations of Howard's model are presented (Tables 1–3). In order to obtain

TABLE 1—L^s (LABOR SUPPLY) EQUATION

	Constant	B^0	H_{-1}/P	W/P	P^A/P^B	
OLS						
Estimated Coefficient	.5592	.001479	6345	- 0204	- 0856	
Point Elasticity[a]	–	6451	0615	–.6560	– 1822	$R^2 = .9820$
t-Ratio	11.4648	5.0787	1 5260	–3.3790	–2.7512	$DW = 1.66$
2SLS						
Estimated Coefficient	.4547	000336	1420	.0094	–.1770	
Point Elasticity	–	1466	.0138	3023	–.3767	$R^2 = .9530$
t-Ratio	3.1479	.3069	.1167	3400	–2.1727	$DW = 1.51$
3SLS						
Estimated Coefficient	.4137	–.00015	0952	.02014	–.1980	
Point Elasticity	–	–.0654	.0092	.6476	–.4214	$R^2 = .9299$
t-Ratio	4.204	–.2210	1706	1.1623	–2.9295	$DW = 1.40$

Note Definitions of variables and data used are as follows: L^s = Annual average nonprivate civilian employment divided by the working-age population (for men 16–59, for women 16–54); B^0 = Official state and cooperative retail sales in 1955 constant prices are divided by the working-age population and then transformed into index (1955 = 100); H_{-1} = Lagged nominal savings (saving deposits plus current accounts plus currency holdings) per working-age population; P = Official consumer price index; P^A = Free market price index; P^B = Official state and cooperative retail price index; W = Average monthy wages; A^d = Free market sales in 1955 prices are divided by the working-age population and then transformed into index (1955 = 100); s^d = The change in real savings (savings deposits plus current accounts plus currency holdings) per working-age population. Most of data used are taken from the *National Statistical Yearbook* (*Rocznik Statystyczny*), 1956–76. Number of observations are same in all estimations ($n = 21$).

[a]Point elasticities are calculated at means of variables.

TABLE 2—A^d (FREE MARKET DEMAND) EQUATION

	Constant	B^0	H_{-1}/P	W/P	P^A/P^B	
OLS						
Estimated Coefficient	339.899	.9322	−79.659	−14.084	−187.759	
Point Elasticity	–	1 7679	−.0336	−1.9691	1.7374	R^2 = .7909
t-Ratio	6.8346	3 1023	−.1879	−2.2289	−5.9197	DW = 1.89
2SLS						
Estimated Coefficient	325 013	3006	−205.312	1.534	−274.450	
Point Elasticity	–	.5701	−.0865	.2145	2.5397	R^2 = .6525
t-Ratio	2 7674	.3376	−.2965	.0684	−4.1439	DW = 1 98
3SLS						
Estimated Coefficient	342.651	.1872	52.669	1.5143	−279 331	
Point Elasticity	–	.3550	.0222	.2117	2.5850	R^2 = .6383
t-Ratio	5.6977	5837	−.1748	.1892	−5.8020	DW = 2.00

Note. See Table 1 for definitions of variables and data used.

results for Poland comparable with Howard's for the Soviet Union, I tried to adhere to his model. Nevertheless the following minor modifications were made:

(i) To eliminate equation (8) from the system, without suppressing information on P^B, the relative price P^A/P^B was used. But it should be noticed that P^A/P^B is not a ratio of the absolute levels of prices on the controlled market and the free market. The latter is more indicative of excess demand on the controlled market. The replacement of two price variables P^A/P and P^B/P by the ratio of two price indices might be generally justified by relatively small influences of price expectation factors on the behavior of agents, due to small movements of prices in *CPE* countries.

(ii) Secondly, in order to identify the free market supply equation, the variables included in this equation were changed. There is no theoretical reason to include an asset variable H_{-1}/P in the commodity supply equation. At the same time, Howard's equation does not include other important factors affecting supply behavior, such as the amount of state purchases of agricultural products or the gross agricultural output, the movement of state procurement prices or their ratio to free market prices, or the money income of peasants. Therefore, as a first attempt H_{-1}/P was dropped from the supply equation and instead, an index of state purchases of agricultural products was included. The real wage W/P was not dropped at the initial stage on the basis of the theoretical plausibility that a

TABLE 3—s^d (SAVINGS DEMAND) EQUATION

	Constant	B^0	H_{-1}/P	W/P	P^A/P^B	
OLS						
Estimated Coefficient	−.0378	− 000078	.0375	0026	.0188	
Point Elasticity	–	−1.9640	2103	4.8513	2.3128	R^2 = .8889
t-Ratio	−2.3410	−.7977	.2723	1.2684	1.8232	DW = 1 02
2SLS						
Estimated Coefficient	−.0401	−.000068	.0311	.0024	.0224	
Point Elasticity	–	−1.7140	.1743	4.4610	2.7557	R^2 = .8880
t-Ratio	−1.3483	−.2995	.1775	.4196	1.3371	DW = 1.05
3SLS						
Estimated Coefficient	−.0426	−.000084	.01830	.00285	.02171	
Point Elasticity	–	−2.1206	.1026	5.3030	2.6708	R^2 = .8880
t-Ratio	−1.4352	−.3734	.1045	.5027	1.2952	DW = 1.03

Note: See Table 1 for definitions of variables and data used.

rise in W/P might lead to movements of the labor force from rural areas to the town, resulting in a reduction in supply of free market goods. However, in order to identify the supply function from the demand, W/P was later dropped. This can be justified by the insignificant t-ratio of the coefficient of W/P. A dummy variable (R), which takes value 1 in case of bad weather and value 0 otherwise, remains as an exogenous variable in the supply equation, but another dummy variable representing government policy (G) is dropped.

(iii) Despite the problems with Howard's model, his specification was largely retained for purposes of comparison; in particular W/P was retained as an endogenous variable. Since data were not available on capital stock, lagged Net Material Product was included instead.

The tables give ordinary least squares (*OLS*), *2SLS*, and *3SLS* estimates for the first three equations. The results of estimating Howard's model applied to the Polish case are not so impressive as Howard's results for the Soviet Union. Further, comparison between estimates obtained by different techniques shows that the results are very sensitive to the estimation techniques used. All simultaneous estimates are inferior to single equation *OLS* estimates according to all statistical criteria. This results largely from the presence of strong multicollinearity between exogenous variables used in this system. But it also indicates problems in the specification of the system. Moreover, comparison of the Polish and Soviet results suggests the model's insensitivity to frequent changes in policy and circumstances such as those in the Polish case. In fact, in view of the nonuniformity and nonhomogeneity of the consumer sector in the Polish economy during the postwar period, it is difficult to impose a unified assumption of disequilibrium in the maintained hypothesis in relatively long periods. Portes and Winter (1978b) suggest that the consumption goods market in Poland was characterized as much by excess supply as by excess demand for the period 1954–75, though the pattern was changed substantially from year to year. The division of the entire period into subperiods would restrict the number of degrees of freedom for estimation. The available observations (twenty-one in this case) also seem relatively insufficient for the application of full-information techniques to simultaneous models of the size under consideration (note that Howard uses only thirteen observations).

REFERENCES

R. J. Barro and H. Grossman, "Suppressed Inflation and the Supply Multiplier," *Rev. Econ. Stud.*, Jan. 1974, *41*, 87–104.

J. Farrell, "Bank Control of the Wage Fund in Poland, 1950–1970," *Soviet Stud.*, Apr. 1975, *27*, 265–87.

F. D. Holzman, "Comment on Soviet Monetary Controls through the Banking System," in Gregory Grossman, ed., *Value and Plan*, Berkeley; Los Angeles 1960.

D. H. Howard, "The Disequilibrium Model in a Controlled Economy: An Empirical Test of the Barro-Grossman Model," *Amer. Econ. Rev.*, Dec. 1976, *66*, 872–79.

B. G. Katz, "A Note on Empirical Tests of the Barro-Grossman Disequilibrium Model in a Controlled Economy," mimeo., New York Univ. 1978.

A. N. D. McAuley, "Wage Determination in the USSR," mimeo., University of London 1976.

G. S. Maddala and F. Nelson, "Maximum Likelihood Methods for Markets in Disequilibrium," *Econometrica*, Nov. 1974, *42*, 1013–30.

M. K. Nissanke, "An Examination of a Suggested Test of Disequilibrium in the Polish Economy (1955–1975)," econ. disc. paper no. 61, Birkbeck College, London, Oct. 1978.

R. D. Portes and D. Winter, "The Supply of Consumption Goods in Centrally Planned Economies," *J. Compar. Econ.*, Dec. 1977, *1*, 351–65.

——— **and** ———, (1978a) "The Demand for Money and for Consumption Goods in Centrally Planned Economies," *Rev. Econ. Statist.*, Feb. 1978, *60*, 8–18.

——— **and** ———, (1978b) "Disequilibrium Estimates for Consumption Goods Markets in Centrally Planned Economies," disc.

paper no. 612, Harvard Inst. Econ. Res., Mar. 1978; *Rev. Econ. Stud.*, forthcoming.

S. **Rudcenko,** "Household Money Income, Expenditure and Monetary Assets in Czechoslovakia, GDR, Hungary and Poland, 1956–1975," in Hans Raupach, ed., *Jahrbuch der Wirtschafts Osteuropas*, Munich 1978, Vol. 8.

Central Statistical Office of Polish People's Republic, *Statistical Yearbook* (*Rocznik Statystyczny*), Warsaw 1975–76.

The Disequilibrium Model in a Controlled Economy: Reply and Further Results

By DAVID H. HOWARD*

Barbara Katz and Machiko Nissanke have presented two comments that are critical of my application of the Barro-Grossman disequilibrium model to the Soviet Union. In the empirical work reported in my 1976 article, I found that the disequilibrium model performed well in that the Soviet data were consistent with the model's predictions. In the present paper, I will not only reply to the criticisms in the Katz and Nissanke papers, but will also report some additional results.

I. Katz

Katz's first criticism has to do with my measure of B^0—goods available on the constrained market. My measure of B^0 is goods sold on the state and cooperative retail markets and is, according to Katz, flawed in two ways. First, Katz states that because B^0 plus saving virtually add up to income, the two components must have a negative relationship, and thus my results are uninteresting. Second, Katz claims that the changing mix of highly desired and inferior goods in B^0 makes it an increasingly unreliable measure of goods availability. This is allegedly due to the fact that as spending on desired goods increases, inventories of inferior goods accumulate despite generalized excess demand.

Katz's second criticism refers to my neglect of any consideration of the effects of the change in political regime that took place during the sample period. Katz introduces a binary variable reflecting the change in regimes and reestimates my equation for household saving. Her variable (D) takes on the same (negative) sign as I found for B^0 and has a t-statistic of nearly 2. She then discards several of the variables in my equation and runs another household saving equation with just the relative price of collective farm market goods and D as explanatory variables. In this regression D has a significant and positive coefficient. She then concludes that the change in political regime explains saving behavior and that the disequilibrium model is thereby rejected. Katz explains her result by stating that D is acting as a proxy for household expectations about consumer good availabilities. Thus, under Brezhnev the households expected greater supplies of consumer goods and saved in order to accumulate funds to purchase them. Katz also points out that my data indicate that savings grew at a much faster rate than did the wage bill or retail sales. From this, she infers that the disequilibrium model's predictions are contradicted.

II. Nissanke

Nissanke begins by stating that I claim that my results support both the hypothesis that there was repressed inflation in the Soviet Union in 1955–67, and the various predictions of the Barro-Grossman disequilibrium model. She next cites two alleged econometric problems in my study. First, the A^s equation is not identified, thereby raising the question of how my results were obtained. Second, according to Nissanke, W (the wage rate) should not be taken as endogenous for a variety of reasons. Apparently these reasons include the lack of justification of my L^d equation, the L^d equation's vertical slope, lack of consideration of planners' behavior, and the fact that W is a government control variable.

Next, Nissanke turns to what are termed more fundamental problems. The first is that my model does not consider planners' behavior and that it is incorrect to build a simultaneous model without including planners'

*International Finance Division, Federal Reserve Board. This paper represents my views solely and should not be interpreted as reflecting the views of the Board of Governors of the Federal Reserve System or other members of its staff.

behavior. Thus, according to Nissanke, I should have used single equation techniques when estimating the equations or I should have modeled planners' behavior. Nissanke also states that my approach assumes the existence of disequilibrium in the B^0 market and equilibrium in the labor market and thus is an inadequate test of the disequilibrium model. Quoting Nissanke, "But for a more elaborate analysis of disequilibrium, the *existence* of disequilibrium on the commodity market itself should be investigated properly. It is the latter with which we are most concerned, and without careful examination of this aspect, a proper interpretation of results is not possible" (p. 728–29). Nissanke reiterates Katz's point about the interpretation of the B^0 coefficients, that is the definition of B^0 ensures that the coefficients always will be what I found them to be. She then returns to my econometric techniques by saying that I should have used disequilibrium estimation techniques, in which a $Q = \min(D,S)$ condition replaces an equilibrium condition.

Finally, Nissanke modifies my model and applies it to Poland. Her modifications are to replace P^A/P and P^B/P with a single price ratio P^A/P^B, to eliminate H_{-1}/P from A^s, and to add and delete a few other variables in A^s. Her empirical results seem to reject the disequilibrium model or her specification of it or the hypothesis of continuous repressed inflation during her sample. Nissanke points out also that her results are sensitive to the estimation technique used. Nissanke explains her results for the Polish economy by citing the different circumstances prevailing in Poland—frequent policy changes and the possibility of switches from excess demand to excess supply conditions during the sample period.

III. Reply

A. *Katz*

With regard to Katz's claim that there is a necessary negative relationship between B^0 and s^d (change in savings bank deposits), it is important to note that the sum $(B^0 + s^d)$ does not equal income. In my model there is one additional market—A^d, the collective farm market. In addition there are several other markets that I did not model due to a lack of data, for example, cash hoarding and black markets. Thus even if one could take income as fixed, the budget constraint in this model does not imply a negative relationship between B^0 and s^d. Katz's general point about the prediction that saving and consumption are inversely related is more properly a criticism of the Barro-Grossman disequilibrium model rather than my implementation of it, but nevertheless the point is still incorrect. As I have discussed elsewhere (1977, pp. 408–09), the budget constraint implies that a reduction in the consumption constraint will lead to increased saving or increased leisure (or some combination of changes in saving and leisure). The analysis in my 1977 article shows that it is optimal for both saving and leisure to increase when consumption decreases if the goods are substitutes. The crucial point here, which was missed by Katz, is that labor supply is a choice variable in the case of repressed inflation and therefore income is not fixed. Furthermore, the disequilibrium model produces predictions about spillovers on all markets. Thus the appropriate way to judge the model is to look at all the B^0 coefficients, not just one in isolation.[1] I have argued in this paragraph that the $\partial s^d/\partial B^0$ prediction is not trivial; however, even if it were, it would not detract from the model unless all of the other implications were also trivial.

Katz also criticizes my measure of B^0, the constrained consumer goods. My measure certainly is not perfect but is the best that I could find. Perhaps Katz's comments will stimulate efforts to find or construct a better measure. However, Katz's point about unsold stocks making sales a bad proxy for availabilities is incorrect. In fact, use of retail sales

[1]Katz considers the s^d response to be the most important. Although my results indicate that the s^d spillover is the most responsive in terms of elasticities, I consider the labor supply response to be the most interesting one. Certainly in terms of the Barro-Grossman "supply multiplier," the labor response is the key spillover effect.

avoids the unsold stocks problem. As I have argued elsewhere:

> An apparent problem results from the use of the data on state and cooperative retail sales as the measure of constrained consumer goods (or the quantity constraint). These data include figures for goods that are not in excess demand. The use of sales, not production, alleviates this by eliminating inventory buildups. It can be shown that using total sales of a group of goods as a constrained good is valid even if some of those goods are not in excess demand. Let x_i represent volume of sales of good i, x_i^d represent demand for it, and p_i its price. If the good is in excess demand then
>
> $$p_i x_i < p_i x_i^d,$$
>
> if not, then by the principle of free consumer choice
>
> $$p_i x_i = p_i x_i^d,$$
>
> i.e., the household need never purchase more than it wants. If there are m goods in the group then
>
> $$p_i x_i \leq p_i x_i^d \qquad i = 1, \ldots, m.$$
>
> If there is just one good in the group for which there is a shortage then
>
> $$\sum_{i=1}^{m} p_i x_i < \sum_{i=1}^{m} p_i x_i^d.$$
>
> Thus, use of $\Sigma_{i=1}^{m} p_i x_i$ as the quantity constraint is valid, since the sum is less than the value of total demand for that group of goods. The group as a whole, then, can be viewed as a constrained consumer good even though consumption of some goods in the group is not constrained. [1979, ch. 6]

Thus, as long as some goods in a particular group are in excess demand, that is, as long as some prices are too low, one can treat the entire group as being in excess demand. A change in the constraint will then cause spillover effects, albeit within the group as well as outside the group. The incorrect prices put the consumer in disequilibrium and the model's predictions hold. These predictions have to do with the substitutability of the goods involved; if the goods in excess demand are substitutes for leisure and future goods (and in the Soviet case A^d goods), then changes in the constraints produce the predicted results, regardless of the existence of some goods in the group that are not in excess demand. However, there is a danger that at some point much of the spillover will be into other B^0 goods—see the discussion of this point in my reply to Nissanke—and in this way the changing mix between inferior goods and desired goods is potentially important.

Katz also criticizes my work for neglecting the change in political regime that took place during my sample period. Katz's first set of empirical results indicate that her D variable is capturing the same effect as my B^0 variable since it has a negative coefficient. Under Brezhnev ($D = 1$), B^0 increased, thus the disequilibrium model would predict $\partial s^d/\partial D < 0$, which is what she found. Katz then deletes some variables, presumably to avoid multicollinearity, and obtains results in which the D coefficient changes signs and the remaining variable from my model (P^A/P) has a t-ratio of less than 2. Given these results it would probably make sense to delete P^A/P and the constant—both of which have t-ratios less than 2—and run the regression again. But then to what would her model be reduced? It would tell us that s^d went up in the Brezhnev years; but we know that already just from looking at the data. It would seem better to fit the model suggested by analysis of economic behavior under disequilibrium conditions and live with the probable large variances caused by multicollinearity, which is what my approach does.

Nevertheless, Katz's point about the effect of expectations is a good one and expectations would work the way she claims. However, the effect is not easily modeled. In particular, why would one model expectations the way Katz does? Is it reasonable that households knew immediately that conditions under Brezhnev would be better? It would seem that some function of present and past B^0 would be a better proxy of expectations than is Katz's binary variable D. There are too few data points to model expectations formation formally, but present B^0 probably is a fairly

good proxy of the determinants of expected B^0. If so, then there are offsetting effects of changes in B^0 on saving and this could explain the relatively weak significance of B^0 in the s^d equation (i.e., weak relative to the B^0 coefficients in the other equations).

Katz's final point has to do with the relative growth of savings during the sample period. However, comparison of the growth in the stock of savings with the growth in the flow of income or consumption is misleading and uninteresting. For example, a constant marginal propensity to save and a constant income imply an increase in the stock of savings and, of course, no change in income or consumption. Joyce Pickersgill has examined data on average and marginal propensities to save in the Soviet Union during 1955–71. She concludes that her "results do not support the view that Soviet households are saving larger and larger proportions of their increased incomes" (p. 146). However, the course of variables such as the average propensity to save (for example) is irrelevant to the question at hand since examination of such variables omits the influence of other factors. The disequilibrium model involves several factors and options operating on and open to the household when faced with a quantity constraint, and proper evaluation of the model involves examination of all of them. For example, increasing B^0 availability is consistent with a rising average propensity to save if wages are rising and saving is sufficiently wage elastic.

Although, as I have indicated above, there are several problems with Katz's comment on my paper, I think that her paper is a useful contribution in that it points out the potential importance of expectations in the analysis. In fact, the influence of expectations may very well be the explanation of the relatively weak significance of B^0 in the s^d equations reported in my paper.

B. *Nissanke*

At the outset I should state that my paper assumes the existence of repressed inflation in the sense that the controlled prices of at least some consumer goods are below market-clearing levels (see below). Therefore my tests are of a joint hypothesis of the presence of this particular kind of repressed inflation and the disequilibrium model. My results do not contradict this joint hypothesis, but they do not constitute a direct test of the existence of repressed inflation in the Soviet Union. Thus, Nissanke's representation of my conclusions in the first paragraph of her comment is somewhat misleading.

As Nissanke points out, the A^s equation is unidentified. In fact, contrary to what Nissanke says, A^s is unidentified by the order condition even after P^B/P is eliminated (P^B/P is eliminated to avoid perfect multicollinearity). The procedure for dealing with an unidentified equation in *3SLS* estimation is to delete that equation from the equations to be estimated (see Arnold Zellner and Henri Theil, p. 68), which is what I did.

Turning to the alleged exogeneity of W, Nissanke's argument is rather unstructured and I will just respond to each point in the order in which she makes them. First, the justification of the demand for labor function in my model is as follows. Let the demand for labor (L^d) depend primarily on output (Q), given the capital stock (K), and treat both Q and K as exogenous since they are primarily policy variables. Next, allow some sensitivity of L^d to the real wage (W/P). Thus L^d (Q, $K,W/P$) would seem to be a reasonable specification, with the expectation that L^d is not very sensitive to W/P given the importance of achieving Q. Note that an L^d that is not very sensitive to W/P means that L^d in Figure 2 of my article is steeply sloped but not that it is completely vertical as Nissanke seems to imply. Finally, it should be pointed out that a slope of this nature was used in my discussion of feedback effects and policy but was not imposed in the estimation procedure. Furthermore, if the real wage adjusts to equate supply and demand on the labor market, the real wage is endogenous regardless of the wage sensitivity of labor demand. Nissanke also criticizes my L^d function for its alleged neglect of planners' decisions; but with both Q and K entering explicitly as exogenous policy

variables (see my 1976 paper, p. 873) it is hard to understand why she makes this criticism. The inclusion of the real wage in L^d is to allow for some ability on the part of enterprises to adjust their demand for labor, but the effect of planners' preferences on enterprise behavior is also modeled by including Q and K. Nissanke's final point about the exogeneity of W concerns the extent to which W is a government policy variable. Certainly the government tries to control wage expenditures in the Soviet Union and has a good deal of influence on them. However, enterprises did have a substantial amount of leeway in practice with regard to wages during my sample period (see Robert Fearn). I did not ignore these considerations (see p. 872, fn. 3, and p. 874) as Nissanke claims; on the contrary, she ignores Fearn's argument, which was cited in my paper.

Certainly it is true that modeling planners' behavior would be a useful extension of my work. But it is not true that a model that considers household behavior alone and treats policy as exogenous—a fairly common procedure—is of no value. It should also be pointed out that some of the informal discussion in my 1976 paper does examine possible policy reactions and feedback effects, although these are not modeled formally. (See my book, ch. 5, for a further discussion of policy reactions and feedback effects, particularly with respect to implications for labor freedom.) Nevertheless, policy modeling is an area for more research.

Next Nissanke criticizes me for assuming disequilibrium in the B^0 market and equilibrium in the labor market. I did make these assumptions but it is important to understand the kind of repressed inflation in the B^0 market that is assumed. All that is needed is for some goods to have prices that are too low for market clearing. These disequilibria create spillovers into leisure, saving, and other substitutes (see my 1977 article for the n-good case). The literature on the Soviet economy indicates that during the sample period used in my study there were some prices of important types of state-supplied consumer goods that were set too low. (The fact that they were important types of goods means that spillovers probably went beyond other state-supplied consumer goods.) Even Richard Portes, who criticizes the literature for failing to distinguish clearly between general repressed inflation and incorrect relative prices, agrees that some relative prices in the Soviet Union are set below market-clearing levels. In the passage quoted above, Nissanke states that she is most concerned with investigating the existence of disequilibrium; however I was not, given the literature attesting to its existence in at least the form of incorrect relative prices. I was most interested in testing the predictions of the disequilibrium model *given* the existence of disequilibrium. On a related matter, I did not use the disequilibrium estimating technique suggested by Nissanke because if, as I have argued above, $S < D$ at each point of time in the sample, then one can just set $Q = S$, which is what I did, in effect.

With regard to Nissanke's argument that my definition of B^0 implies that my results are trivial, see above my reply to Katz's similar point.

Next, I would like to discuss Nissanke's modification of my model and her empirical results for Poland. I have no real objection to putting P^A/P^B in place of P^A/P and P^B/P, although it appears to be just an *ad hoc* change. Nissanke then ends up with seven equations and eight unknowns unless she treats W as exogenous (in the spirit of her comments on my paper). Nissanke's deletion of real balances (H_{-1}/P) from A^s is not permissible however. There is a very good theoretical reason for including H_{-1}/P in A^s: the household sector is doing a great deal of the supplying in this market (the collective farm market) and the household sector's functions should include its assets, the relative prices of goods it buys and sells, and the quantity constraints it faces, as well as any other relevant variables.

Turning to Nissanke's empirical results, it is interesting that she—like Katz—argues that my definition of B^0 means that the coefficient estimates that I reported should always be obtained, but Nissanke's results—

like Katz's—contradict their argument. Apparently the predictions are not so trivial after all! (See above my reply to Katz on this point.) Nissanke points out that her empirical results are sensitive to the estimation technique used. In my 1979 study (ch. 7, Appendix), I present some results using *OLS* and *2SLS*. These results indicate that in the Soviet case, the conclusions are not sensitive to the estimating technique used. Nissanke's explanation of her results for Poland seems reasonable. Nissanke discusses the potential importance of modeling planners' behavior explicitly and her results for Poland may very well be evidence supporting this point.

IV. Further Results

Regressions reported in my book (ch. 7, Appendix) indicate that the results reported in my 1976 article (p. 875, Tables 2 and 3) are not sensitive to the estimation technique used. In particular, the sizes of the coefficients of B^0 as well as their t-ratios are approximately the same across estimation methods. In addition, the sizes of the t-ratios of the B^0 coefficients in the s^d equation—the equation discussed by Katz—are larger in the *2SLS* and *OLS* regressions than they are in the regressions reported in my 1976 paper. In my 1979 study (ch. 7) several different data and model specifications are used to test the robustness of the results reported in my 1976 article. The conclusion of this investigation is that the results are very robust. In particular, the crucial B^0 coefficients do not change signs and with one exception their t-ratios exceed 2 in absolute value, and the exception has a t-ratio of 1.98. (For further details about the results referred to in this section, see my book, ch. 7.)

REFERENCES

R. J. Barro and H. I. Grossman, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 83–93.

——— **and** ———, "Suppressed Inflation and the Supply Multiplier," *Rev. Econ. Stud.*, Jan. 1974, *41*, 97–104.

R. M. Fearn, "Controls Over Wage Funds and Inflationary Pressures in the U.S.S.R.," *Ind. Labor Rel. Rev.*, Jan. 1965, *18*, 186–95.

David H. Howard, "The Disequilibrium Model in a Controlled Economy: An Empirical Test of the Barro-Grossman Model," *Amer. Econ. Rev.*, Dec. 1976, *66*, 871–79.

———, "Rationing, Quantity Constraints, and Consumption Theory," *Econometrica*, Mar. 1977, *45*, 399–412.

———, *The Disequilibrium Model in a Controlled Economy*, Lexington 1979.

B. G. Katz, "The Disequilibrium Model in a Controlled Economy: Comment," *Amer Econ. Rev.*, Sept. 1979, *69*, 721–25.

M. K. Nissanke, "The Disequilibrium Model in a Controlled Economy: Comment," *Amer Econ. Rev.*, Sept. 1979, *69*, 726–32.

J. Pickersgill, "Soviet Household Saving Behavior," *Rev. Econ. Statist.*, May 1976, *53*, 139–47.

R. Portes, "Macroeconomic Equilibrium Under Central Planning," seminar paper no. 40, Inst. Int. Econ. Stud., Stockholm, Sept 1974.

A. Zellner and H. Theil, "Three-Stage Least Squares: Simultaneous Estimation of Simultaneous Equations," *Econometrica*, Jan. 1962, *30*, 54–78.

On the Estimation of Disaggregate Welfare Losses with an Application to Price Distortions in Urban Transport

By STEPHEN GLAISTER*

Abram Bergson (1973) has questioned the proposition that welfare losses due to monopoly-induced deviation from efficient prices are quantitatively unimportant. His paper attracted comments from Richard Carson and Dean Worcester to which he made his reply (1975). His argument was essentially that one must avoid the measurement errors and biases caused by using broad aggregate categories of goods and by ignoring the variation of price-marginal cost differentials within each aggregate. His demonstration was principally a numerical analysis of one special case. Christopher Foster has summarized the importance of Bergson's contribution in its bearing on the work of Richard Pryke and others on the performance of the *U.K.* nationalized industries. The first aim of this paper is to clarify and generalize the theoretical principles underlying Bergson's results and place them in the context of conventional public finance theory. The second aim is to illustrate the resulting calculus with a case study of a public enterprise which, because of externally imposed public policy, has been forced to hold its prices below the relevant marginal costs.

Bergson showed that welfare losses will depend upon the own-price elasticities of the goods in question and upon the variability of the price distortions, in addition to their magnitudes. It is shown below that they will also depend upon the cross elasticities between the goods. Bergson's formulation involved strong restrictions on these. David L. Lewis and I have had the opportunity of estimating demand elasticities for London Transport's bus and underground rail services, and evidence was found for substantial cross elasticities between these two substitute services. The period was one during which money fares were held relatively constant so that fares in real terms fell very substantially—by about one-third between January 1972 and March 1975. I apply the method developed to this data to see if (a) the welfare loss, "correctly" calculated, due to this policy of price restraint could be said to be significant and (b) how closely calculations at various levels of aggregation, which failed to take account of the variations in price distortions or of the cross elasticities between the component parts of the aggregate, might approximate to the "correct" value.

I. A Generalization and Extention of Bergson's Results

The aim of this section is to indicate that Bergson's discussion of welfare losses in the general equilibrium context can be formulated concisely in terms of the standard consumer's expenditure function. His numerical example is a special case of a more general formula.

Following Bergson, assume that there are n commodities whose total quantities in the economy are represented by the vector $x = (x_1, x_2, \dots, x_n)$. It is convenient in this context to assume that they can be produced at constant marginal costs, which would also be the competitive prices $p^c = (p_1^c, p_2^c, \dots, p_n^c)$. It is also convenient to regard the consuming community as if it were a single individual.

Peter Diamond and Daniel McFadden and my 1974 paper discuss the properties and uses of the individual's expenditure function. It is denoted by $g(p, u)$ and represents the minimum amount of money required for the consumer to achieve utility level u at prices p. A useful property is that its partial derivative with respect to a particular price is the

*The London School of Economics. I am indebted to Christopher Foster and a referee for advice in preparing this paper.

compensated demand function for that good (see Diamond and McFadden, p.4):

$$(1) \quad \frac{\partial g(p,u)}{\partial p_i} = x_i(p,u) \qquad i = 1, 2, \ldots, n$$

Consider the expression

$$(2) \quad \{g(p^m,u) - g(p^c,u)\} - \{p^m x^m - p^c x^m\}$$

where x^m is a vector whose elements are $x_i(p^m,u)$. The utility level u is held constant throughout expression (2).

The first term in brackets is the compensating variation for a price change from the competitive to the distorted prices; what Bergson calls the gross compensating variation (*GCV*). Since p^c measures the average unit production costs, the second term in brackets measures the profit that would be earned by changing prices from p^c to p^m if incomes were changed so as to hold attainable utility constant at the competitive equilibrium level (Bergson's *MIA*). Under the constant cost assumptions this term accounts for the fact that in the case of a price increase the move from x^c to x^m would release resources which would be reemployed in a full-employment general equilibrium.

Expression (2) is therefore the net compensating variation (*NCV*). Bearing in mind that by definition

$$g(p^m,u) = p^m x^m$$

and dividing (2) by expenditure at competitive prices we may rewrite it as

$$(3) \qquad L = \frac{p^c x(p^m,u) - g(p^c,u)}{g(p^c,u)}$$

This corresponds to Diamond and McFadden's expression for a welfare loss due to an excise tax and it gives precisely the measure used by Bergson, *CNCV*, for his special case. This may be checked by deriving the expenditure and compensated demand functions for his constant elasticity of substitution utility function, and substituting the results in (3). This immediately yields the expression which he used for his numerical evaluation.

The expression in (3) is a definition of a welfare loss measure, and the formulae Bergson used were exact special cases of it. However, such special cases are always restrictive and they are often intractable. Thus it is convenient to have an alternative, but approximate, expression for (3). Taking the first three terms in the Taylor expansion of $g(p^c,u)$ in the numerator of (3) about the point p^m yields

$$Lg(p^c,u) \simeq p^c x(p^m,u) - g(p^m,u) - \sum_i \frac{\partial g(p^m,u)}{\partial p_i}(p_i^c - p_i^m) - \frac{1}{2}\sum_i\sum_j \frac{\partial^2 g(p^m,u)}{\partial p_i \partial p_j}(p_i^c - p_i^m)(p_j^c - p_j^m)$$

$$= -\frac{1}{2}\sum_i\sum_j \frac{\partial x_i(p^m,u)}{\partial p_j}(p_i^c - p_i^m)(p_j^c - p_j^m)$$

by virtue of (1).

Letting

$$(4) \qquad dp_i = p_i^m - p_i^c$$

be the ith price distortion we have

$$(5) \qquad Lg(p^c,u) \simeq -\frac{1}{2}\sum_i\sum_j \frac{\partial x_i}{\partial p_j} dp_i\, dp_j$$

a fundamental relationship in indirect taxation theory. This must be positive for any set of dp_i—positive or negative—because the pure substitution matrix is negative definite. Equivalently, if we define

$$\rho_i = \frac{p_i^m - p_i^c}{p_i^c}$$

as the ith proportionate price distortion and

$$\eta_{ij} = \frac{\partial x_i}{\partial p_j} \cdot \frac{p_j}{x_i}$$

as the elasticity of demand for good i with respect to the price of good j, then we have

$$(6) \qquad L \simeq -\frac{1}{2}\sum_i\sum_j \gamma_i \eta_{ij} \rho_i \rho_j$$

where

$$\gamma_i = \frac{p_i x_i}{g(p^c,u)}$$

is the fraction of income spent on the ith good. The approximation in (5) or (6) will be good so long as the price distortions (4) are small.

Equations (5) and (6) show that the welfare losses will depend in a quite general fashion on the precise way in which the own-price elasticities, price distortions, and expenditure shares interact—as Bergson demonstrated. They show that the loss equally involves the corresponding cross elasticities between goods, a fact which Bergson's approach did not emphasize since his particular utility function enforces the strong restriction that $\eta_{ii} = \gamma_i(\sigma - 1)$, where σ is the elasticity of substitution.

Given that one is prepared to accept the approximation involved and that one can remove income effects, it is relatively easy to apply (5) and (6) directly to a calculation of welfare loss. The general formula is rather complicated and it is useful to write out the special case where only goods 1 and 2 are distorted, all other prices remaining at their competitive levels. From (6)

$$(7) \quad -2L \simeq \gamma_1(\eta_{11}\rho_1^2 + \eta_{12}\rho_1\rho_2) + \gamma_2(\eta_{21}\rho_1\rho_2 + \eta_{22}\rho_2^2) = (\gamma_1\eta_{11}\rho_1^2 + \gamma_2\eta_{22}\rho_2^2) + \rho_1\rho_2(\gamma_1\eta_{12} + \gamma_2\eta_{21})$$

The first term in parentheses here is necessarily negative. If both price distortions are in the same direction and the goods are substitutes then the second term will be positive. Therefore its omission would cause (7) to overestimate the loss. On the other hand if the price distortions are in opposite directions then the omission of the cross-elasticity terms would lead to an underestimate of the loss.

II. Errors Due to Aggregation

This brings us to the question of the likely nature of the errors made by using invalid measures. Bergson made calculations in two cases. In one he assumed there was one monopolized product taking half the expenditure. In the other he assumed there were two monopolized products each taking one-quarter of the expenditure, and with the same weighted-average price distortion as the first case. The second case showed consistently higher welfare loss than the first. The difference between these results could be interpreted as indicating the error that would be made by assuming that two distinct monopolized goods could be aggregated and treated as if they formed one aggregate good. It is a little difficult to generalize precisely about this, because the error made depends upon the particular method of aggregation. I record here the results for two specific cases, both simplified by assuming that only goods 1 and 2 suffer a price distortion while all others are competitively priced.

First, suppose that while recognizing that goods 1 and 2 are distinct, it is assumed that they both exhibit the same degree of price distortion ρ, which is a weighted average of the actual distortions ρ_1 and ρ_2:

$$\rho = \frac{\gamma_1\rho_1 + \gamma_2\rho_2}{\gamma_1 + \gamma_2}$$

where $\rho_1 < \rho < \rho_2$.

Define L^* to be the welfare loss measure derived from (7) on the assumption that the two goods have the same degree of price distortion. Since the degree of distortion in fact differs, there will be an error in the true value L. This will be given by

$$(8) \quad 2(L - L^*) \simeq \gamma_1\eta_{11}(\rho^2 - \rho_1^2) + \gamma_1\eta_{12}(\rho^2 - \rho_1\rho_2) + \gamma_2\eta_{21}(\rho^2 - \rho_1\rho_2) + \gamma_2\eta_{22}(\rho^2 - \rho_2^2)$$

This expresses twice the error (approximately) in terms of the two own-price elasticities and the cross elasticities. The coefficients of these elasticities involve the proportions of income spent on the respective goods, and the deviations of the price changes about their weighted-average change. That these factors should be relevant seems reasonable and is confirmed by Bergson's calculations.

I now turn to a consideration of the likely sign and magnitude of the error. Since this is essentially the difference between two negative definite quadratic forms nothing can be said in general, but some results can be obtained in special cases. First note that, as

one would expect, if the two prices are raised by the same proportion so that $\rho_1 = \rho_2 = \rho$, then the error is zero. Second, consider the own-price elasticity terms. These may be written

$$\begin{aligned}&\gamma_1\eta_{11}(\rho^2 - \rho_1^2) + \gamma_2\eta_{22}(\rho^2 - \rho_2^2)\\&\quad = \gamma_1\eta_{11}(\rho - \rho_1)(\rho + \rho_1)\\&\qquad + \gamma_2\eta_{22}(\rho - \rho_2)(\rho + \rho_2)\\&\quad = \gamma_1(\rho - \rho_1)[\eta_{11}(\rho + \rho_1) - \eta_{22}(\rho + \rho_2)]\end{aligned}$$

If the two goods are of a very similar nature it is reasonable to assume that they exhibit similar compensated own-price elasticities. If so then it is likely that the expression will be positive, since the elasticities are negative. However, it should be noted that this conclusion might be reversed if the larger price deviation is associated with the smaller (in absolute value) price elasticity. This might well be the case in practice if the price deviations are the result of monopolistic pricing, but none of Bergson's examples have this property—rather the reverse, guaranteeing the positivity of the expression. Consider the remaining expression in (8)

$$(\gamma_1\eta_{12} + \gamma_2\eta_{21})(\rho^2 - \rho_1\rho_2)$$

Assuming the two goods to be substitutes, the cross elasticities will be positive and the sign will be the same as that of the terms in the second pair of parentheses. This in turn depends upon the relationship between two numbers and their weighted average. It can be shown that

$$(\rho^2 - \rho_1\rho_2) \geq 0$$

$$\text{if} \qquad \frac{\gamma_1}{\gamma_1 + \gamma_2} \leq \frac{1}{1 + \sqrt{\rho_1}/\sqrt{\rho_2}}$$

In other words, since the right-hand side of this inequality is larger than one-half, the whole of the expression involving the cross elasticities will be positive unless the expenditure on the good with the lower price distortion is a significantly greater proportion of income than that on the good with the higher distortion. In particular, if the expenditures are roughly equal (or if the good with the greater distortion also accounts for a higher proportion of expenditure), then the expression will be positive. It is interesting to note that each of Bergson's "asymmetrical" cases in his second table is exactly on the borderline, so that only the own-price effects remain. There the less-distorted good receives twice the expenditure of the other one.

The second special case we consider is as follows. Suppose that goods 1 and 2 are incorrectly lumped together and treated as one good, and that its price elasticity η is estimated from expenditure data. Then we would have

$$\eta = \frac{\gamma_1(\eta_{11} + \eta_{12}) + \gamma_2(\eta_{21} + \eta_{22})}{(\gamma_1 + \gamma_2)}$$

The natural aggregate price distortion to use would be ρ as defined above. The expression for welfare loss would then be

$$(9) \qquad L^* \simeq -\tfrac{1}{2}(\gamma_1 + \gamma_2)\eta\rho^2$$

Hence the error would be given by

$$\begin{aligned}2(L - L^*) \simeq {} & [\gamma_1(\eta_{11} + \eta_{12})\\&+ \gamma_2(\eta_{21} + \eta_{22})]\left[\frac{\gamma_1\rho_1 + \gamma_2\rho_2}{(\gamma_1 + \gamma_2)}\right]^2\\&- \gamma_1\eta_{11}\rho_1^2 - \gamma_2\eta_{22}\rho_2^2\\&- \gamma_1\eta_{12}\rho_1\rho_2 - \gamma_2\eta_{21}\rho_2\rho_1\end{aligned}$$

Some manipulation converts this expression into the equivalent form

$$\begin{aligned}(10) \quad 2(L - L^*) \simeq {} & \frac{\gamma_1\gamma_2(\rho_1 - \rho_2)}{(\gamma_1 + \gamma_2)}\\&\cdot\left[\eta_{22}(\rho + \rho_2) - \eta_{11}(\rho + \rho_1)\right.\\&\left.+ \frac{(\gamma_1\eta_{12} + \gamma_2\eta_{21})}{(\gamma_1 + \gamma_2)}\frac{(\gamma_1^2\rho_1 - \gamma_2^2\rho_2)}{\gamma_1\gamma_2}\right]\end{aligned}$$

Equation (10) demonstrates that the aggregation error in this case depends upon a weighted average of the cross elasticities as well as upon the own-price elasticities.

As in the previous cases it is not possible to generalize about the sign of (10) and hence about the direction of the error. However, a sufficient condition for the aggregate calculation to produce an underestimate of the true

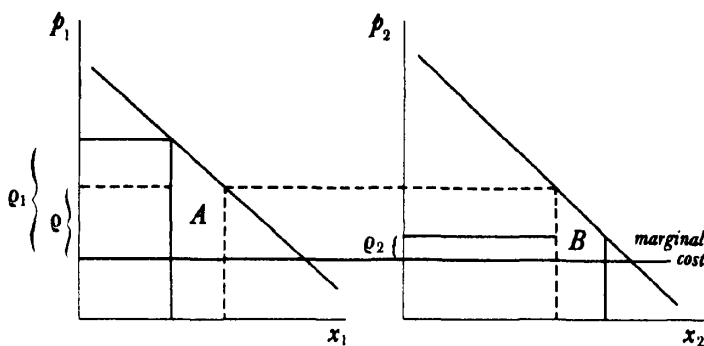



FIGURE 1

welfare loss is that $\gamma_1 < \gamma_2$ and $|\eta_{11}| < |\eta_{22}|$, assuming as before that $\rho_1 < \rho_2$. This confirms the result obtained from (8): that aggregation will lead to underestimation unless expenditure on the good with the lower price distortion is a significantly greater proportion of income than that on the good with the higher distortion, so long as the former has the lower elasticity.

There is a simple graphical interpretation[1] of these results in the special case where both industries have the same own-price elasticity of demand, both command equal expenditure shares, both have the same (constant) marginal costs and both are assumed to have the same markup rates although they actually have different markup rates. Figure 1 shows that using the average markup ρ understates the true loss in market 1 by an amount A, while it overstates it by B in market 2; A is larger than B by an amount which depends on the common elasticity of demand.

After calculating the implicit own and cross elasticities in Bergson's (1973) formulation, equation (10) reproduces the differences between his Tables 1 and 2 quite accurately, given the approximation involved in deriving it, except in those cases where his elasticity of demand (through the elasticity of substitution) is unreasonably large and the price distortions are very big. Hence for moderate demand elasticities and price distortions (10) may be regarded as an analytical generalization of Bergson's results. To give an example with $\sigma = 4$, $\rho_1 = 0.1$, $\rho_2 = 0.3$, and $\gamma_1 = \gamma_2 = 0.25$ the difference between Bergson's Table 1 (col. 3) and Table 2 (col. 3) is (2.43 − 1.66) = 0.77 percent; equation (10) gives it as 1.0 percent. If $\sigma = 16$ Bergson's difference is (9.82 − 5.47) = 4.35, while (10) gives 4.0 percent. If $\sigma = 16$, $\rho_1 = 0.15$, $\rho_2 = 0.60$, and $\gamma_1 = 0.33$ and $\gamma_2 = 0.17$, Bergson's difference is (27.08 − 9.95) = 17.13 percent, while (10) gives 18.21 percent. Dropping the term in the cross elasticities in (10) significantly affects the results, illustrating that the correct specification of these cross elasticities is vital.

III. An Application to London Transport Services

The purpose of this section is to illustrate the theory developed above by application to London Transport's two substitute services, bus and rail. Data on elasticities and costs have been collected for other purposes in the paper by Lewis and myself. The questions to be investigated here are (a) what is the magnitude of the welfare loss due to the failure to adopt optimum pricing for London's bus and underground rail services? (b) How much was this increased by price restraint over the last few years, which considerably depressed fare levels in real terms below what they would otherwise have been? (c) What would be the magnitude of the error in answering questions (a) and (b) when treating bus and rail services as a single aggregate good? The data used in this section are discussed at length in the paper by Lewis and myself; because of the uncertainties about the

[1] I am indebted to a referee for this interpretation.

most appropriate values the calculations carried out in this section should only be regarded as illustrative of the foregoing sections.

London underground fares were increased in the thirty-seventh week of 1972 and then both bus and underground fares remained unchanged until the twelfth week of 1975. During this time the retail price index rose from 100.0 to 144.1 and an index of earnings rose somewhat more than this. Therefore real fares fell substantially during a period when, other things being equal, they should have been increased slightly. On the assumption that the fares were at their optimum levels at the beginning of this period we now use (6) to estimate the rate of welfare loss associated with this period of price restraint.

The *Family Expenditure Survey* shows that in 1972 total expenditure in Greater London at current prices was £1.05 million per week, of which about 1.12 percent was spent on bus services and 1.11 percent on London Transport rail. Table 1 shows a set of own- and cross-fare compensated elasticities derived by aggregation from Table 2, which is explained below.

Substituting these values in (6) with $\rho_1 = \rho_2 = -0.306$ gives a welfare loss of 0.02 percent. At 1972 prices this represents about 0.7 pence per household per week, or a total of £1.1 million per year for the whole Greater London area. This is about 0.9 percent of total revenue from the two modes. Although this is a small quantity it comes from distortions in the prices of commodities which form a very small proportion of total expenditure. This is because London Transport's services form a small proportion of total transport expenditure in the area, which is about 13 percent of all expenditure on average. If this exercise were to be repeated for the whole of the transport sector—or indeed for all components of expenditure—then the welfare losses might be substantial. However, one cannot escape the conclusion that the losses due to the price "distortion," as measured here, in London Transport's fares due to price restraint over this period was not large. In fact, failure to maintain real wages contributed to a staff shortage which caused a serious deterioration in service quality. This may well have caused a much more significant welfare loss. Traffic congestion externalities are ignored in this calculation and they too may be significant. There is no question of aggregation bias in this case since the two price distortions are the same.

TABLE 1

	Demand for	
	Bus	Rail
Bus Fare	−0.43	0.15
Rail Fare	0.16	−0.34

During 1975 there were two fare increases which succeeded in restoring real bus fares to roughly what they had been at the beginning of 1970 and increasing real underground fares somewhat above previous levels. In 1976 both fares were about 4.2p. per passenger mile on average. On the other hand, Lewis and I calculated an "optimum" set of peak and off-peak bus and rail fares when marginal social costs and cross elasticities between private and public transport are taken into account. The set of compensated elasticities, derived from various sources is shown in Table 2. The table also shows the "optimum" fares, the proportionate distortions of actual 1976 fares from these optima ρ, and the predicted shares of expenditure at the optimum γ.

With these parameter values formula (6) gives an estimated welfare loss of 2.99 percent. If, however, peak and off-peak data are aggregated in the manner described in the previous section to give just two services, bus and rail, then we obtain proportionate price distortions of − 0.27 and 0.36, respectively, expenditure shares 0.0086 and 0.019 and the set of elasticities shown in Table 1. Equation (6) then estimates the welfare loss at 0.0765 percent.

Aggregating further to give just one aggregate public transport service gives a price distortion of 0.16, an expenditure share of 0.027, and an elasticity of − 0.22. Equation (6) then estimates the welfare loss at 0.00796 percent. Equation (10) confirms the differ-

TABLE 2

Elasticities with respect to:		Bus Peak	Bus Off-Peak	Rail Peak	Rail Off-Peak
Bus Fare	Peak	−0.35	0.04	0.14	0.01
	Off-Peak	0.029	−0.87	0.009	0.28
Rail Fare	Peak	0.143	0.013	−0.30	0.05
	Off-Peak	0.008	0.28	0.018	−0.75
Optimum fares (pence)		7.02	3.12	6.11	0.54
ρ		−0.40	0.35	−0.31	6.78
γ		0.0072	0.0015	0.017	0.0018

ence between the "two-mode" and the "one-mode" cases as 0.0765 − 0.00796 = 0.06854 percent.

The purpose of this exercise is simply to illustrate how aggregation of commodities which should truly be regarded as separate can lead to underestimation of welfare losses. I have shown how disaggregating a single service to two services and then to four services increases estimated losses in the proportions 1: 9.6: 39.1. However, one caveat should be entered here: equation (6) and its derivatives are Taylor approximations to the "true" loss. In the fully disaggregate case some of the price distortions may be sufficiently large to cast doubt on the accuracy of this approximation. The averaging procedure implicit in the aggregation makes this less of a problem for the more aggregate cases.

IV. Summary and Conclusions

I have attempted to provide an explanation and approximate analytical generalization of Bergson's (1973) results. I have shown that in estimating welfare losses due to price distortions it is important to work at the most disaggregate level possible and to take full account of cross-price elasticities between goods and services. Failure to do this will in general lead to errors which will often (but not necessarily always) be underestimates of true welfare losses.

The results have been illustrated by application to crude data for London Transport's bus and underground services from 1970–75. Welfare losses due to price restraint over the period appeared to be small as a proportion of all expenditure, although they are greater as a proportion of expenditure on the services themselves. Those due to failures in 1976 fares to match optimum marginal social cost fares are much greater, especially if it is reckoned that peak services should be distinguished from off-peak services. Severe underestimation of welfare losses results from aggregate calculations.

The numerical calculations presented here are really only intended as illustrations of the analytical results. As propositions about London Transport's pricing policy they must be treated with reserve. There are many difficulties but I mention two. First, the assumption that the "base" prices are indeed the optimum prices is debatable, especially in the case of the price-restraint argument. Second, some of the price distortions are so large that the approximation involved may not be valid. To be exact one would need to know the forms of the demand functions, as Bergson did by assumption, and to work out explicit forms for the welfare losses.

In summary this paper supports the use of the measure proposed by Bergson (1973) for welfare losses due to price distortion, attempts to throw some light on its properties, and supports the proposition that accurate measurement requires a disaggregate treatment. An illustrative example has demonstrated cases where welfare losses are not negligible and aggregation leads to serious underestimation of them.

REFERENCES

A. Bergson, "On Monoply Welfare Losses," *Amer. Econ. Rev.*, Dec. 1973, *63*, 853–70.

———, "On Monopoly Welfare Losses: Reply," *Amer. Econ. Rev.*, Dec. 1975, *65*, 1024–31.

R. Carson, "On Monopoly Welfare Losses: Comment," *Amer. Econ. Rev.*, Dec. 1975, *65*, 1008–14.

P. Diamond and D. McFadden, "Some Uses of the Expenditure Function in Public Finance," *J. Publ. Econ.*, Feb. 1974, *3*, 3–22.

Christopher D. Foster, *From Policy to Administration*, London 1976.

S. Glaister, "Generalised Consumer Surplus and Public Transport Pricing," *Econ. J.*, Dec. 1974, *84*, 849–67.

——— **and D. Lewis,** "An Integrated Fares Policy for Transport in London," *J. Publ. Econ.*, June 1978, *9*, 341–55.

Richard Pryke, *Public Enterprise in Practice*, London 1971.

D. A. Worcester, "On Monopoly Welfare Losses: Comment," *Amer. Econ. Rev.*, Dec. 1975, *65*, 1015–23.

U.K. Department of Employment, *Family Expenditure Survey, Report for 1972*, London 1973.

NOTES

NINETY-SECOND ANNUAL MEETING OF THE AMERICAN ECONOMIC ASSOCIATION

Atlanta, Georgia, December 27–30, 1979

Preliminary Announcement of the Program

Thursday, December 27, 1979

10 00 A.M. EXECUTIVE COMMITTEE MEETING

Friday, December 28, 1979

8.00 A.M. CHANGING ROLE OF THE MARKET IN UTILITIES REGULATION* (Joint Session with the Transportation and Public Utilities Group of the AEA)
Presiding: WILLIAM MELODY, Simon Fraser University
Papers: HARRY M. TREBING, Michigan State University
Structural Change and Regulatory Reform in the Utilities Industries
LELAND K. JOHNSON, National Telecommunications and Information Administration
Marketing a Social Resource: The Radio Spectrum
THOMAS K. STANDISH, Connecticut Public Utilities Control Authority
State Initiatives and State/Federal Relations
Discussant: DALLAS W. SMYTH, Simon Fraser University

8 00 A.M. OMICRON DELTA EPSILON INVITED STUDENT PAPERS (Joint Session with Omicron Delta Epsilon)
Presiding: G. RANDOLPH RICE, Louisiana State University

8 00 A.M. INTERGOVERNMENTAL AID AND THE FINANCE OF EDUCATION
Presiding: ELCHANAN COHN, University of South Carolina
Papers: JAY CHAMBERS AND HENRY M. LEVIN, Stanford University
The Welfare Implications of Proposition 13
ANITA A. SUMMERS, Federal Reserve Bank of Philadelphia, AND BARBARA L. WOLFE, University of Wisconsin
Improving the Use of Empirical Research as a Policy Tool: An Application to Education
JOHN RIEW, Pennsylvania State University
Enrollment Decline and Cost-Benefit Analysis of School Reorganization
ELCHANAN COHN AND JAMES R. SWEIGART, University of South Carolina
A New Approach to Financing Public Schools
Discussants: CHARLES S. BENSON, University of California-Berkeley
O. H. BROWNLEE, University of Minnesota
W. LEE HANSEN, University of Wisconsin
WALTER W. MCMAHON, University of Illinois

8 00 A.M. THE INCIDENCE AND EFFECTIVENESS OF GOVERNMENT TAX AND TRANSFER PROGRAMS
Presiding: HAYNES C. GODDARD, University of Cincinnati
Papers: ROBERT A. MOFFITT, Rutgers University
A Review of the Recent Literature on the Disincentive Effects of Tax Rates
RICHARD V. BURKHAUSER AND TIMOTHY M. SMEEDING, Institute for Research on Poverty, University of Wisconsin
OASI, SSI, and Low Earners
MARC BENDICK, JR. AND JAMES P. ZAIS, The Urban Institute
What Should be the Role of Demand-Side Subsidies in U.S. Housing Policy: Lessons from Experiments with Housing Allowances
Discussants: CHARLES P. WALDAUER, Widener College
MICHAEL L. GOETZ, Temple University
DOUGLAS Y. THORSEN, Bradley University

8:00 A.M. THREE ASPECTS OF ECONOMIC DEVELOPMENT
Presiding: JEFFREY B. NUGENT, University of Southern California
Papers: GARY S FIELDS, Cornell University
Poverty, Equality, and Development: A Distributional Approach
ANDRÉ SAPIR, University of Wisconsin-Madison
Economic Growth and Factor Substitution: What Happened to the Yugoslav Miracle?
SITHA BABU, Rider College
Economic Analysis of Fertility in Less Developed Countries
Discussants: JAMES H GAPINSKI, Florida State University
NICHOLAS SANCHEZ, College of the Holy Cross
ROGER R. BETANCOURT, University of Maryland

10:15 A.M. THE GOVERNMENT AND CAPITAL FORMATION*
Presiding: GEORGE VON FURSTENBERG, International Monetary Fund
Papers. ROBERT E. HALL, Stanford University
Long-Run Affects of Demand Management Policies
MICHAEL J. BOSKIN AND JOHN B SHOVEN, Stanford University
The U.S. Tax System and Incentives to Save and Invest: Where Do We Stand?
Discussants: ALAN BLINDER, Princeton University
HARVEY GALPER, U.S Department of the Treasury

10:15 A.M. INTERWAR MACROECONOMICS FROM THE PERSPECTIVE OF THE 1970'S* (To mark the Fiftieth Anniversary of the Onset of the Great Depression)
Presiding. ROBERT J GORDON, Northwestern University
Papers: JACOB FRENKEL, University of Chicago
Prices, Money, and Exchange Rates: Lessons from the 1920's
ROBERT J. GORDON, Northwestern University
The Behavior of Prices in the Interwar and Postwar Periods: Is There a Consistent Explanation?
CHRISTOPHER A. SIMS, University of Minnesota
Comparing Interwar and Postwar Business Cycles
Discussants. ALLAN MELTZER, Carnegie-Mellon University
JOEL MOKYR, Northwestern University
JEFFREY SACHS, National Bureau of Economic Research

10:15 A.M THE EFFECTS OF FISCAL POLICIES ON THE DISTRIBUTION OF INCOME AND WEALTH*
Presiding RICHARD A MUSGRAVE, Harvard University and University of California-Santa Cruz
Papers: MERVYN KING, University of Birmingham
How Effective Have Fiscal Policies Been in Changing the Distribution of Income and Wealth?
RICHARD BIRD, University of Toronto
Methodological Problems in Assessing the Distributional Effects of Fiscal Policies
LESTER THUROW, Massachusetts Institute of Technology
Interactions Between Fiscal Policies and Other Policy Instruments Affecting Distribution
Discussants. RICHARD A MUSGRAVE, Harvard University and University of California-Santa Cruz
JAMES MEDOFF, Harvard University

10:15 A.M. INNOVATION, TECHNOLOGICAL PROGRESS, AND RESEARCH AND DEVELOPMENT*
Presiding: NATHAN ROSENBERG, Stanford University
Papers: GEORGE EADS, Rand Corporation
Impact of Government Regulatory Activities Upon Innovation
NESTOR TERLECKYJ, National Planning Association
What R & D Data Tell Us About the Determinants of Technical Change
RICHARD R. NELSON, Yale University
Interindustry Comparisons of Technical Change
Discussant: ROLF PIEKARZ, National Science Foundation

10:15 A.M. WOMEN'S PLACE IN THE LABOR MARKET: WILL ERA MATTER? (Joint Session with the AEA Committee on the Status of Women in the Economics Profession)
Presiding: ANN F. FRIEDLAENDER, Massachusetts Institute of Technology
Papers: HEIDI I. HARTMANN, National Academy of Science
Comparable Worth and the ERA
JAMES J. HECKMAN, University of Chicago and National Bureau of Economic Research
The Likely Impact of ERA on Women's Economic Status

FRANK P. STAFFORD, University of Michigan
Prospects for Women's Earnings: The Roles of Policy and the Market
Discussants: LAWRENCE M. KAHN, University of Illinois-Urbana
MICHAEL REICH, University of California-Berkeley
MYRA H. STROBER, Graduate School of Business, Stanford University

10.15 A.M. BLACK ECONOMIC PROGRESS: THE REALITY AND THE ILLUSION (Joint Session with the National Economic Association)
Presiding: GLENN LOURY, Northwestern University
Papers: CHARLES BETSY, Congressional Budget Office
Black-White Unemployment Rates Through Time
LINDA DATCHER, University of Michigan
Relative Growth of Black Income Since the 1960's
DONG K. JEONG, North Carolina Agriculture and Technical State University
A Study of Change in Impacts of Discrimination on Black/White Earnings Differentials During 1960–70: A Simultaneous-Equations Approach
WILLIAM DARITY AND SAMUEL MYERS, University of Texas
Black-White Inequality over the Business Cycle
Discussants: GLENN LOURY, Northwestern University
RICHARD FREEMAN, Harvard University
JAMES SMITH, Rand Corporation

10.15 A.M. WORLD DEVELOPMENT IN THE 1980'S: THE ISSUES AND THEIR MODELING (Joint Session with the Society for Policy Modeling)
Presiding: DOUGLAS O. WALKER, United Nations
Papers: F. G. ADAMS, University of Pennsylvania
The Law of the Sea and Mineral Policy: An Econometric Approach to Evaluating the Impact of Manganese Nodal Production
A. M. COSTA, United Nations and New York University
Alternative Modeling Procedures for the Study of Development and Restructuring the World Economy
S. GUPTA AND J. KREUSER, World Bank
The Competitive vs. the General Equilibrium Approach. An Exercise in Global Modeling at the World Bank for Medium-Term Policy Making
S. M. MENSHIKOV, United Nations
Modeling for International Policy Making: The Case of the U.N. International Development Strategy for the 1980's
Discussants: V. FILATOV, United Nations and Queens College
S. LAWSON, Fordham University

10.15 A.M. GOVERNMENT REGULATION IN THE HEALTH CARE FIELD (Joint Session with the Health Economics Research Organization)
Presiding: DONALD E. YETT, University of Southern California
Papers: WAYNE WENDLING, Education Commission of the States, AND JACK WERNER, American Medical Association
The Economic Theory of Regulation: An Application to Hospitals
RONALD J. VOGEL AND NANCY T. GREENSPAN, Health Care Financing Administration
An Econometric Analysis of the Effects of Regulation on the Private Health Insurance Market
PAUL B. GINSBURG AND LARRY WILSON, Congressional Budget Office
A Model for Analyzing Hospital Cost Containment Proposals
Discussants: HENRY J. AARON, The Brookings Institution
WILLIAM LYNK, Blue Cross and Blue Shield Associations
MICHAEL D. INTRILIGATOR, University of California-Los Angeles and University of Southern California

10.15 A.M. ECONOMICS OF HOUSING
Presiding. HUGH NOURSE, College of Business Administration, University of Georgia
Papers: JAMES A. VERBRUGGE, University of Georgia, AND THOMAS J. PARLIAMENT, U.S. League of Savings Associations
Tax-Exempt Bonds for Single-Family Housing: An Evaluation of State House Finance Agency and Local Government Programs
LLOYD A. TURNER, University of Iowa
Alternative Mortgage Instruments and the Local Housing Market

JOHN M. CLAPP, University of California-Los Angeles
Appraisal Practices as a Cause of Mortgage Deficiency
Discussants: A. JAMES HEINS, University of Illinois
JAMES LITTLE, Washington University (St Louis)
A. THOMAS KING, Federal Home Loan Bank Board

10:15 A.M THE ECONOMICS OF PREDATORY PRICING
Presiding, ALAN R BECKENSTEIN, University of Virginia
Papers: PAUL JOSKOW, Massachusetts Institute of Technology, AND ALVIN KLEVORICK, Yale University
The Bases of Alternative Predatory Pricing Rules
ROLAND H KOLLER, Brigham Young University
Predatory Pricing The Tip of the Iceberg
H. LANDIS GABEL AND ALAN R BECKENSTEIN, University of Virginia
The Economics of Intent and Other Qualifications to the Debate on Predatory Pricing
ROBERT WILLIG, Princeton University, AND J A ORDOVER, New York University
An Economic Definition of Predation
Discussants PETER O STEINER, University of Michigan
KENNETH ELZINGA, University of Virginia
RICHARD MARKOVITS, University of Texas Law School
F M SCHERER, Northwestern University

10:15 A M STABILIZATION POLICIES IN INTERDEPENDENT ECONOMIES
Presiding WILLIAM P. SNAVELY, George Mason University
Papers LEONARD DUDLEY, University of Montreal
Bilateral Trade Flows under Endogenous Stabilization Policy
WILLEM H BUITER AND JONATHAN EATON, Woodrow Wilson School, Princeton University
Decentralized Policies for Stabilization and Growth
CARL VAN DUYNE, Institute for International Economic Studies, University of Stockholm, and Williams College
Are Export Controls Anti-Inflationary?
DAVID BIGMAN, Hebrew University, AND PENELOPE SMITH, International Monetary Fund
The Dollar Crisis A Conflict between Theory and Evidence
Discussant MAXWELL J FRY, University of Hawaii

10.15 A M. THE ROLE OF NATURAL RESOURCES IN ECONOMIC PROCESSES (Joint Session with the Association of Environmental and Resource Economists)
Presiding V KERRY SMITH, University of North Carolina
Papers: TJALLING KOOPMANS AND WILLIAM NORDHAUS, Yale University
(Title to be announced)
ROBERT GORDON AND BRIAN SKINNER, Yale University
(Title to be announced)
GEOFFREY HEAL, University of Sussex and Columbia University
(Title to be announced)

12:30 P.M. JOINT LUNCHEON WITH THE AMERICAN FINANCE ASSOCIATION
Presiding EDWARD J KANE, Ohio State University
Speaker JOHN G HEIMANN, U S Comptroller of the Currency

2:00 P.M BORDERLINES OF LAW AND ECONOMIC THEORY*
Presiding ROBERT CLOWER, University of California-Los Angeles
Papers. BENJAMIN KLEIN, University of California-Los Angeles
Contract Law and Economic Theory
WESLEY J LIEBELER, University of California-Los Angeles
Sixty-Eight Volumes of the *AER*
A. MITCHELL POLINSKY, Harvard University
Strict Liability vs Negligence in a Market Setting
Discussants: STEVEN N. CHEUNG, University of Washington
PAUL L JOSKOW, Massachusetts Institute of Technology
PATRICIA MUNCH, Rand Corporation

2:00 P.M. THE CURRENT RETARDATION IN U.S PRODUCTIVITY GROWTH*
Presiding: SOLOMON FABRICANT, National Bureau of Economic Research and New York University

Papers: ALBERT REES, Alfred P. Sloan Foundation
Improving Productivity Measurements
ZVI GRILICHES, Harvard University
R & D and the Productivity-Growth Slowdown
M. ISHAQ NADIRI, New York University
Postwar Behavior of Sectoral Productivity in the United States
Discussants: JEROME A. MARK, U.S. Bureau of Labor Statistics
EDWARD F. DENISON, Department of Commerce
RICHARD R. NELSON, Yale University

2:00 P.M. RECENT DEVELOPMENTS IN THE ECONOMIC THEORY OF INDEX NUMBERS*
Presiding: ROBERT M. SOLOW, Massachusetts Institute of Technology
Papers. W. ERWIN DIEWERT, University of British Columbia
Capital and the Theory of Productivity Measurement
DALE W. JORGENSON, Harvard University, LAWRENCE J. LAU, Stanford University, AND THOMAS M. STOKER, Sloan School, Massachusetts Institute of Technology
Welfare Comparison under Exact Aggregation
ROBERT A POLLAK, University of Pennsylvania
Recent Developments in the Theory of Cost-of-Living Index Numbers
Discussants. LAURITS R. CHRISTENSEN, University of Wisconsin
JOHN S CHIPMAN, University of Minnesota
JACK E TRIPLETT, U.S Bureau of Labor Statistics

2:00 P.M. TAXATION, LABOR SUPPLY, AND SAVING*
Presiding: RUDOLPH G. PENNER, American Enterprise Institute
Papers: HOWARD ROSEN, Princeton University
Labor Supply
GEORGE M VON FURSTENBERG, International Monetary Fund
Private Saving
MARTIN S FELDSTEIN, National Bureau of Economic Research and Harvard University
Policy Implications
Discussants MARVIN KOSTERS, American Enterprise Institute
JOHN SHOVEN, Stanford University
ALAN S BLINDER, Princeton University

2:00 P.M. APPRAISALS OF POST-KEYNESIAN ECONOMICS*
Presiding WALTER S. SALANT, The Brookings Institution
Papers: LORIE TARSHIS, Scarborough College, University of Toronto
(Title to be announced)
JANET YELLEN, London School of Economics
(Title to be announced)
JAMES CROTTY, University of Massachusetts
(Title to be announced)
Discussants: GEOFFREY C. HARCOURT, University of Adelaide
PAUL WELLS, University of Illinois-Champaign

2:00 P.M. NEW METHODS IN MONEY DEMAND ANALYSIS (Joint Session with the Econometric Society)
Presiding: JOHN TAYLOR, Columbia University
Papers: BENJAMIN FRIEDMAN, Harvard University
Denison's Law and the Relative Stability of Money and Credit "Velocities"
DONALD HESTER, University of Wisconsin
Multivariate Analysis of a Panel of Weekly Reporting Banks: Evidence of Technical Change in Intermediation
PETER TINSLEY AND BONNIE GARRETT, Board of Governors of the Federal Reserve System
Measurement of Money Demand
Discussants: JOHN SCADDING, Federal Reserve Bank of San Francisco
RAY FAIR, Yale University
MICHAEL HAMBURGER, New York University

2:00 P.M. THE IDEA OF ECONOMIC EXPLOITATION: THREE PERSPECTIVES (Joint Session with the Association for Evolutionary Economics and the Association for Social Economics)
Presiding: LANE VANDERSLICE, University of Notre Dame

Papers: HERBERT GINTIS, University of Massachusetts
Exploitation and Domination. On the Absence of Concepts in Neoclassical Welfare Theory
WARREN J. SAMUELS, Michigan State University
A Critique of the Concept of Exploitation in the History of Economics
STEPHEN T. WORLAND, University of Notre Dame
What Religion Has to Tell Economics About Exploitation
Discussants: JAMES BASS, InterAmerican Development Bank
PHILIP A. NEHER, University of British Columbia
MARTIN BRONFENBRENNER, Duke University

2:00 P.M. WORLDWIDE REDISTRIBUTION OF INCOME
Presiding: IRMA ADELMAN, University of California-Berkeley
Papers: (To be announced)

2:00 P.M. ECONOMIC ASPECTS OF ENVIRONMENTAL REGULATIONS
Presiding: E. H. WARREN, JR., Jet Propulsion Laboratory, California Institute of Technology
Papers. ANTHONY C. FISHER, University of California-Berkeley
Integrated Assessment for Environmental Standards: How Well Has it Been Done?
A. MYRICK FREEMAN, Bowdoin College
Empirical Estimates of the Benefits of Controlling Air and Water Pollution: A Review and Synthesis
DANIEL T. DICK AND EDNA T. LOEHMAN, SRI International
Benefit-Cost Analysis of Hazardous Chemicals. A Case Study of Fluorocarbon Control
TIMOTHY H. QUINN, ADELE R. PALMER, AND GEORGE C. EADS, Rand Corporation
Marketable Permits as a Strategy for Pollution Control
Discussants: ROGER G. NOLL, California Institute of Technology
E. H. WARREN, JR., Jet Propulsion Laboratory

2:00 P.M. EMPIRICAL STUDIES OF PRODUCTIVITY GROWTH
Presiding: MIEKO NISHIMIZU, Princeton University
Papers: MICHAEL ELLIS, North Texas State University
A Time-Series Analysis of Intrasectoral Changes in Capacity Utilization, Productivity, and Unit Labor Costs
BADIUL A. MAJUMDAR, Albers School of Business, Seattle University
The Empirical Relevance of the Theory of Endogenous Technical Change
SUSAN A. SIMMONS, University of Southwestern Louisiana, S. CABELL SHULL, AND MICKEY C. SMITH, University of Mississippi
Rates of Return on Research and Development Expenditures in the U.S. and U.K. Pharmaceutical Industries
Discussants: PEARL M. KAMER, New York State Department of Labor
BABU NAHATA, University of Louisville

2:30 P.M. THE BUSINESS OUTLOOK FOR 1980 (Joint Session with the American Finance Association)
Presiding: BURTON G. MALKIEL, Princeton University
Speakers. WILLIAM A. COX, U.S. Department of Commerce
OTTO ECKSTEIN, Harvard University and Data Resources, Inc.
MICHAEL K. EVANS, Evans Economics, Inc
ALAN GREENSPAN, Townsend-Greenspan & Co., Inc.
LAWRENCE R. KLEIN, University of Pennsylvania and Wharton Econometric Forecasting Associates
HAROLD T. SHAPIRO, University of Michigan
ALBERT M. WOJNILOWER, First Boston Corporation

8:00 P.M. RICHARD T. ELY LECTURE*
Presiding: MOSES ABRAMOVITZ, Stanford University
Speaker: TIBOR SCITOVSKY, Stanford University
Can Capitalism Survive?—An Old Question in a New Setting

Saturday, December 29, 1979

8:00 A.M. IMPLICATIONS OF DEREGULATION (Joint Session with the Transportation and Public Utilities Group of the AEA)
Presiding: WILLIAM G. SHEPHERD, University of Michigan
Papers: WALTER G. BOLTER, Federal Communications Commission
Communications

TED KEELER, University of California-Berkeley
Airlines
DAVID S. SCHWARTZ, Public Interest Economist, Washington, D.C.
Natural Gas Production
Discussant: NATHANIEL B. CLARKE, Interstate Commerce Commission

8 00 A.M. COMPENSATION TO THOSE HARMED BY PUBLIC ACTIONS
Presiding: JOSEPH J. CORDES, George Washington University
Papers: JOSEPH J. CORDES, George Washington University, AND BURTON A. WEISBROD, University of Wisconsin-Madison
Compensation to those Harmed by Public Actions: Answered and Unanswered Questions
ROBERT S. GOLDFARB, George Washington University
Severance Pay Compensation: Alternate Views of Equity and the Airline Deregulation Act
TOM BAYARD AND JIM ORR, Bureau of International Labor Affairs, U.S. Department of Labor
Transitional Equity and Efficiency: An Analysis of U.S. Trade Adjustment Assistance Policies
LAWRENCE BACOW, MICHAEL O'HARE, AND DEBRA SANDERSON, Massachusetts Institute of Technology
Compensation and the Energy Facility Siting Process
Discussants: MANCUR OLSON, University of Maryland
ROBERT HARRIS, The Urban Institute
JOHN H MUTTI, University of Wyoming

8 00 A M CRITERIA FOR FACULTY SALARY ADJUSTMENTS
Presiding ALBERT E. REES, Alfred P Sloan Foundation
Papers. RUDOPH BLITZ AND ANTHONY M TANG, Vanderbilt University
Merit Raises and Academic Tenure Under Inflation
PETER O STEINER AND ROBERT HOLBROOK, University of Michigan
Coping with Academic Salaries in a Period of Adversity
Discussants: MARCUS ALEXIS, Northwestern University
BARBARA B. REAGAN, Southern Methodist University

8 00 A M. EXPERIMENTAL METHODS APPLIED TO MANPOWER SUPPORTED WORK
Presiding. ROBINSON G HOLLISTER, Swarthmore College and Mathematica Policy Research, Inc.
Papers. REBECCA MAYNARD, Mathematica Policy Research, Inc.
Youth A Test of Direct Employment
KATHY DICKINSON AND IRVING PILIAVIN, Institute for Research on Poverty, University of Wisconsin
Ex-Addicts and Ex-Offenders. Employment and Recidivism
STANLEY MASTERS AND IRWIN GARFINKEL, Institute for Research on Poverty, University of Wisconsin
Women from AFDC Work and Welfare
PETER KEMPER, DAVID LONG, AND CRAIG THORNTON, Mathematica Policy Research, Inc.
A Benefit-Cost Analysis of the Supported Work Experiment

8 00 A.M. RATIONAL EXPECTATIONS, PRICES, AND INFLATION CONTROL
Presiding: WILLIAM J FRAZER, Jr , University of Florida
Papers. STEPHEN MCCAFFERTY, Ohio State University
Rational Expectations, Disequilibrium Quantities, and Policy Effectiveness in a Non-Market-Clearing Framework
WILLEM H. BUITER, Woodrow Wilson School, Princeton University
A Critical Evaluation of the New Classical Macroeconomics
JAGDISH HANDA, McGill University
Rational Expectations and Stabilization Policy
SHLOMO MAITAL AND YOEL BENJAMIN, Princeton University
Inflation as Prisoner's Dilemma
Discussants F OWEN IRVINE, JR , Wesleyan University
J. T. BRIMER, Washington State University
ELMER G. WEINS, University of Pennsylvania
RICHARD G. SHEEHAN, James Madison University

8 00 A M TRADE, MONEY, AND INFLATION IN DEVELOPING COUNTRIES
Presiding: PAN YOTOPOULOS, Food Research Institute, Stanford University
Papers: JAMES T. H. TSAO, Mechanicsburg, Pa.
Factor Endowment, Foreign Trade and Economic Development

C. JEVONS LEE, Knox College
Interest-Rate and Short-Term Capital Movement in Developing Economies
JAE WAN CHUNG, George Mason University
Inflation in Two Rapidly Developing Countries: Korea and Taiwan
Discussants: M. SAEED AKHTAR, U.N. Development Program, Pakistan
HOSSEIN ASKARI, Business School, University of Texas-Austin
ICHIRO OTANI, International Monetary Fund
C. GLEZAKUS, California State University-Long Beach

10:15 A.M. THE DECLINE OF AMERICAN CITIES: SELF-CORRECTING AND SELF-AGGRAVATING FORCES*
Presiding: ANTHONY DOWNS, The Brookings Institution
Papers. WILLIAM ALONSO, Population Center, Harvard University
The Effects of Regional Population Movements
KATHERINE BRADBURY, ANTHONY DOWNS, AND KENNETH SMALL, The Brookings Institution
The Dynamics of Central City–Suburban Relations
MICHAEL STEDMAN, City and Regional Planning Department, University of North Carolina
The Processes of Neighborhood Change
Discussants: (To be announced)

10:15 A.M. THE POLITICAL ECONOMY OF NATIONAL HEALTH INSURANCE*
Presiding: VICTOR FUCHS, Stanford University and National Bureau of Economic Research
Papers ALAIN ENTHOVEN, Graduate School of Business, Stanford University
How Interested Groups Have Responded to a Proposal for Competition in Health Services
UWE REINHARDT, Princeton University
Health Insurance and Cost-Containment Policies: The Experience Abroad
Discussants: MARK PAULY, Northwestern University
RICHARD ZECKHAUSER, Harvard University

10:15 A.M. CHANGES IN TRADE SHARES AND ECONOMIC GROWTH*
Presiding. IRVING KRAVIS, University of Pennsylvania
Papers: HOLLIS B CHENERY, World Bank
Interactions Between Industrialization and Exports
ANNE O KRUEGER, University of Minnesota
Trade Policy as an Input to Development
CHARLES P KINDLEBERGER, Massachusetts Institute of Technology
Government Policies and Changing National Shares in World Trade
Discussants. RONALD FINDLAY, Columbia University
CARLOS F. DIAZ-ALEJANDRO, Yale University
GARY HUFBAUER, Treasury Department

10:15 A M LONG-TERM GROWTH IN THIRD-WORLD ECONOMIES*
Presiding JAMES C. INGRAM, University of North Carolina
Papers LLOYD G REYNOLDS, Yale University
Economic Development in Historical Perspective
ALAN HESTON, University of Pennsylvania
The Case of India
ALBERT C FISHLOW, Yale University
The Case of Brazil
Discussants: LUCY CARDWELL, University of Massachusetts
SAMUEL MORLEY, Vanderbilt University
MORRIS D MORRIS, University of Washington

10:15 a.m. EMPIRICAL STUDIES OF THE RATE OF RETURN TO CAPITAL* (Joint Session with the American Finance Association)
Presiding: JOHN B. SHOVEN, Stanford University and National Bureau of Economic Research
Papers: DANIEL M. HOLLAND AND STEWART MYERS, Sloan School of Management, Massachusetts Institute of Technology
Profitability and Capital Costs in Corporate Manufacturing
BARBARA FRAUMENI, Boston College, AND DALE W. JORGENSON, Harvard University
Interindustry Comparisons of Rates of Return to Capital
ARNOLD HARBERGER, University of Chicago
International Comparisons of Rates of Return to Capital

Discussants: BURTON MALKIEL, Princeton University and National Bureau of Economic Research
GEORGE STIGLER, University of Chicago
MARTIN S. FELDSTEIN, Harvard University and National Bureau of Economic Research

10 15 A M CONSEQUENCES OF THE GROWTH OF THE TWO-EARNER FAMILY* (Joint Session with the AEA Committee on the Status of Women in the Economics Profession)
Presiding: JANICE F. MADDEN, University of Pennsylvania
Papers: JANICE F. MADDEN, University of Pennsylvania
Consequences of the Growth in Two-Earner Families in Urban Housing Markets
JULIE A. MATTHAEI, Wellesley College
Consequences of the Growth of the Two-Earner Family: A Breakdown of the Sexual Division of Labor
EDWARD P. LAZEAR, University of Chicago and National Bureau of Economic Research, AND ROBERT MICHAEL, Stanford University and National Bureau of Economic Research
Real Income Equivalence among One- and Two-Earner Families
DANIEL QUINLAN AND JEAN SHACKELFORD, Bucknell University
Labor Force Participation Rates of Women and the Rise of the Two-Earner Family
Discussants: MYRA STROBER, Stanford University
FRANCINE BLAU, University of Illinois

10 15 A M THE FEDERAL RESERVE AUTHORITIES AND THEIR PUBLIC RESPONSIBILITY (Joint Session with the American Finance Association)
Presiding: KARL BRUNNER, University of Rochester and Universität Bern
Panelists: MARK WILLES, Federal Reserve Bank of Minneapolis
JERRY JORDAN, Pittsburgh National Bank
ROBERT AUERBACH, Committee on Currency and Banking of the U.S. House of Representatives
ROBERT RASCHE, Michigan State University

10 15 A M HISTORY OF ECONOMIC THOUGHT (Joint Session with the History of Economics Society)
Presiding: MARK PERLMAN, University of Pittsburgh
Papers: SUSAN K. HOWSON, University of Toronto-Scarborough College
On Hawtrey's Centennial
CARL G. UHR, University of California-Riverside
On Heckscher's Centennial
DONALD F GORDON, City University of New York-Baruch College
On Examining von Thünen's Unpublished Monetary Theory
Discussants: FRANK SPRENG, Brescia College
JOHN LETICHE, University of California-Berkeley
ARTHUR LEIGH, Reed College

10 15 A M THE INSTITUTIONAL ROOTS OF INFLATION (Joint Session with the Association for Study of the Grants Economy)
Presiding: TIBOR SCITOVSKY, Stanford University
Papers: KENNETH E. BOULDING, University of Colorado
Wrong Grants to Wrong People: The Source of Inflation
ABBA P. LERNER, Florida State University
Anti-Inflation Incentives
JANOS HORVATH, Holcomb Research Institute, Butler University
Towards an Institutional Theory of Inflation
Discussants: RICHARD D. BARTEL, U.S. Congress Joint Economic Committee
ALAN A. BROWN, University of Windsor
W. JAMES HERMAN, University of South Florida

10:15 A M THE USES AND LIMITS OF MANPOWER POLICY
Presiding: HOWARD ROSEN, Office of Research and Development, U.S. Department of Labor
Papers: ELI GINZBERG, Columbia University and National Commission for Employment Policy
The $80 Billion Innovation
BERNARD ANDERSON, Rockefeller Foundation
How Much Help did Manpower Programs Provide Minorities and Youth?
JOHN PALMER, The Brookings Institution and U.S. Department of Health, Education, and Welfare
Jobs or Income Transfers
Discussants: MICHAEL BORUS, Ohio State University
JODY ALLEN, U.S. Department of Labor

12:30 P.M. LUNCHEON HONORING THE 1978 NOBEL LAUREATE, HERBERT A. SIMON

2:00 P.M. REGULATORY INTERVENTION IN HISTORICAL PERSPECTIVE*
Presiding. PAUL A. DAVID, Stanford University
Papers PETER TEMIN, Massachusetts Institute of Technology
Regulation in the Drug Industry, 1906–62
THOMAS ULEN, University of Illinois-Urbana
The ICC and the Railroad Industry from 1887 to 1920
JOHN PANZAR, Bell Laboratories
The Civil Aeronautics Board, 1930–70
Discussants (To be announced)

2:00 P.M. A GENERAL VIEW OF CAPITAL FORMATION AND ECONOMIC GROWTH*
Presiding: FRITZ MACHLUP, New York University
Papers: ROBERT M. COEN, Northwestern University, AND BERT G. HICKMAN, Stanford University
Investment and Growth in an Econometric Model of the United States
EDWARD F. DENISON, Bureau of Economic Analysis, U.S. Department of Commerce
Capital's Contribution to Growth and Its Retardation
ROBERT EISNER, Northwestern University
Total Income, Total Investment and Growth
Discussants: OLI HAWRYLYSHYN, George Washington University and Statistics Canada
JOHN W. KENDRICK, George Washington University
FRANK DE LEEUW, Bureau of Economic Analysis, U.S. Department of Commerce

2:00 P.M. THE EVOLVING WORLD DOLLAR STANDARD*
Presiding. CHARLES KINDLEBERGER, Massachusetts Institute of Technology
Papers STEPHEN P. MAGEE, University of Texas-Austin
The Dollar vs. Other National Monies in International Trade
JACOB FRENKEL, University of Chicago
The Efficiency of the Foreign Exchange Market and Measures of Turbulence
RONALD MCKINNON, Stanford University
Dollar Stabilization and American Monetary Policy
Discussants ROBERT J. HODRICK, Graduate School of Industrial Administration, Carnegie-Mellon University
PAUL KRUGMAN, Yale University
JOHN CUDDINGTON, Stanford University

2:00 P.M. THE RECOIL FROM WELFARE CAPITALISM: POLITICAL AND SOCIOLOGICAL PERSPECTIVES*
Presiding ROBERT O. KEOHANE, Stanford University
Papers ALBERT O. HIRSCHMAN, Institute for Advanced Study, Princeton University
The Changing Appreciation of Government-Supplied Goods and Services. Some Structural Considerations
IRA KATZNELSON, University of Chicago
Capitalism, Democracy, and the Origins of Social Policy
Discussant ROBERT KEOHANE, Stanford University

2:00 P.M. GOVERNMENT TRANSPORTATION POLICY IN AN INFLATIONARY ECONOMY (Joint Session with the Transportation and Public Utility Group of the AEA)
Presiding. STANLEY J. HILLE, Kent State University
Papers. JOHN W. FULLER, National Transportation Policy Study Commission
Public Expenditure Policy and Inflation in Transportation
ERNEST R. OLSON, Interstate Commerce Commission
ICC Regulation: A Control of, or Contribution to, Inflation
JAMES R. NELSON, Amherst College
How Does One Handle Monetary Inflation in an Industry of Economic Deflation?
Discussant: JOHN F. DUE, University of Illinois

2:00 P.M. EQUALITY INCENTIVES AND ECONOMIC POLICY* (Joint Session with the National Economic Association)
Presiding: MARGARET C. SIMMS, Atlanta University
Papers: RANSFORD W. PALMER, Howard University
Equality Incentives and Economic Policy
MICHAEL REICH, University of California-Berkeley
The Persistence of Racial Inequality in Urban Areas

STEPHANE WILSON AND ERNST W. STROMSDORFER, ABT Associates
The Work Equity Project: A Labor Supply Study of the Welfare Recipients in a Work and Training Program
LASCELLES ANDERSON, Center for Study for Education, Harvard University
Education and Unequal Exchange: An Exposition
Discussants: MARVIN M. SMITH, The Brookings Institution
HENRY FELDER, SRI International
WILLIE R. TAYLOR, Howard University

2:00 P.M. FEDERAL DEREGULATION AND REGULATORY REFORM (Joint Session with the Econometric Society)
Presiding: DARIUS GASKINS, U.S. Department of Energy
Papers: ELIZABETH BAILEY, Civil Aeronautics Board
Deregulation at the CAB
ROBERT CRANDALL, The Brookings Institution
Health and Safety Regulation
MARCUS ALEXIS, Northwestern University
Reform Efforts at the ICC
Discussant GEORGE EADS, Council of Economic Advisers

2:00 P.M. CHINESE AGRICULTURE DEVELOPMENT, PRODUCTION, AND TRADE (Joint Session with the American Agricultural Economics Association)
Presiding CHARLES Y LIU, U.S Department of Agriculture
Papers ROBERT F. DERNBERGER, Center for Chinese Studies, University of Michigan
Agriculture Development. The Key Link in China's Farm Modernization Program
ANTHONY M TANG, Vanderbilt University
Trend, Policy Cycle, and Weather Disturbance in Chinese Agriculture Production, 1952–78
FREDERIC M SURLS, U S Department of Agriculture
Changing Patterns in Chinese Agricultural Trade
Discussants. JAMES A. KILPATRICK, Central Intelligence Agency
NICHOLAS R LARDY, Yale University

2:00 P.M. THE INTERNATIONAL OIL PRICING SYSTEM (Joint Session with the International Association of Energy Economists)
Presiding. MORRIS A ADELMAN, Massachusetts Institute of Technology
Papers (To be announced)

2:00 P.M. REDISTRIBUTIVE IMPACT OF GOVERNMENTAL REGULATIONS (Joint Session with the Association for the Study of the Grants Economy)
Presiding. JANOS HORVATH, Butler University
Papers WILLIAM S VICKREY, Columbia University
Life Line Rates. Succor or Snare
WILLIAM BAUMOL, Princeton University
On Two Distributive Biases of Price Regulation
MILTON Z KAFOGLIS, University of South Florida
The Transfer of Monopoly Profit to Nonregulated Factors
Discussants. BERNERD BOGAR, Indiana University and Purdue University
DOUGLAS N. ROSS, U S. Congress Joint Economic Committee
ROBERT WILLIG, Princeton University

2:00 P.M. LABOR MARKET STUDIES
Presiding: ROBERT T MICHAEL, National Bureau of Economic Research and Stanford University
Papers: MARY KAY PLANTES, University of Wisconsin
Youth Unemployment and Long-Run Earnings Potential
NORMAN WALZER AND STUART DORSEY, Western Illinois University
Risk, Compensating Differentials, and Liability Rules
LAURIE BASSI, Princeton University
The Substitution Effect of Public Service Employment: Most Recent Evidence
Discussants: TEH-WEI HU, Pennsylvania State University
ROBERT M. HUTCHENS, New York State School of Industrial and Labor Relations
ANDREA BELLER, Radcliffe College

8:00 P.M. PRESIDENTIAL ADDRESS
Presiding: E. CARY BROWN, Massachusetts Institute of Technology
Speaker: ROBERT M. SOLOW, Massachusetts Institute of Technology

9:30 P.M. BUSINESS MEETING

Sunday, December 30, 1979

8:00 A.M. GOVERNMENT PROMOTION OF RAILROAD TRANSPORTATION IN THE 1980's (Joint Session with the Transportation and Public Utility Group of the AEA)
Presiding: DONALD V. HARPER, University of Minnesota
Papers: WILLIAM K. SMITH, General Mills, Inc., and U.S. Railway Association
Should Federal Planning and Funding for the Railroads in the 1980's include the 1970's USRA–Conrail Concept?
JOHN W. INGRAM, Chicago, Rock Island, and Pacific R.R.
Government and the Midwest Railroads in the 1980's
JOHN W. HAZARD, Michigan State University
Government Railroading: The Outlook
Discussant: GRANT M. DAVIS, University of Arkansas

8:00 A.M. ECONOMIC EDUCATION RESEARCH IN ATTITUDES AND PERCEPTIONS (Joint Session with the Joint Council on Economic Education)
Presiding: JEFF R. CLARK, Joint Council on Economic Education
Papers: CALVIN A. KENT AND THOMAS KELLEY, Baylor University
Attitudes and Economic Education: Does it Matter?
KIM SOSIN AND CAMPBELL R. MCCONNELL, University of Nebraska-Lincoln
Student Perceptions on Income Distribution
WILLIAM B. WALSTAD, University of Missouri-St. Louis
Determining the Relationship Between Achievement and Attitude in Economics Learning
Discussants: HOWARD P. TUCKMAN, Florida State University
GEORGE G. DAWSON, Empire State University
FRANK D. TINARI, Seton Hall University

8:00 A.M. TRENDS IN INTRODUCTORY ECONOMICS (Joint Session with the Omicron Delta Epsilon)
Presiding: LOUISE NELSON, Davidson College
Papers: (To be announced)

8:00 A.M. THE ASIAN/AMERICAN EMPLOYMENT PROFILE
Presiding: KAZUO SATO, State University of New York-Buffalo
Papers: MANORANJAN DUTTA, Rutgers University
The Stanford Workshop on the Asian/Pacific American Employment Profile—Issues and Nonissues
YUAN-LI WU, University of San Francisco
The Economic Progress of Chinese Americans Since World War II: Occupation, Employment, Income, and Economic Behavior
MAMORU ISHIKAWA, U.S. Department of Labor
Discrimination in Employment of Asian Americans
KANTA MARWAH, Carleton University
Towards Employment Earnings Profile of the Tri-Minority Professionals: Non-Native, Asian, Female
KI-TAEK CHUN, U.S. Commission on Civil Rights
Job-Related Aspirations and Attainments: An Overview
Discussants: ROBERTO S. MARIANO, University of Pennsylvania
JOSEPH S. CHUNG, Illinois Institute of Technology
GREGORY N. T. HUNG, Howard University
THOMAS O. GILSON, Hawaii University
KAZUO SATO, State University of New York-Buffalo

10:15 A.M. STUDIES OF TEACHING AND LEARNING IN ECONOMICS*
Presiding: W. LEE HANSEN, University of Wisconsin–Madison
Papers: JOHN J. SIEGFRIED AND GEORGE H. SWEENEY, Vanderbilt University
Bias in Economic Education Research from Random and Voluntary Selection into Experimental and Control Groups

WILLIAM E. BECKER, JR. AND MATHEW J. MOREY, Indiana University
Pooled Cross-Section–Time-Series Education Evaluation: Source and Result of, and Correction for, Serially Correlated Errors
MICHAEL K. SALEMI, University of North Carolina, AND GEORGE TAUCHEN, Duke University
Guessing and the Error Structure of Learning Models
Discussant: GARY E. CHAMBERLAIN, University of Wisconsin

10:15 A.M. THE CENTENNIAL OF HENRY GEORGE'S *PROGRESS AND POVERTY* (Joint Session with the History of Economics Society)
Presiding: MASON GAFFNEY, University of California-Riverside
Papers: (To be announced)

10:15 A.M. TOPICS IN APPLIED PRICE THEORY
Presiding: RONALD R. BRAEUTIGAM, Northwestern University
Papers: E. K. CHOI AND S. Y. WU, University of Missouri-Columbia
Wealth, Price, and Commodity Risk: Aversions and Price Flexibility
HARVEY E. LAPAN, Iowa University, AND DOUGLAS M. BROWN, National Bureau of Economic Research and Georgetown University
Long-Run Competitive Equilibrium with Endogenous Factor Prices
MICHAEL SATTINGER, Miami University, Ohio
The Opportunity Cost of Factor Pricing
JOHN M. FINKELSTEIN AND RAYMOND F. GORMAN, Graduate School of Business, Indiana University
A General Equilibrium Framework for the Existence of Financial Intermediaries
Discussants: JOHN SORRENTINO, Temple University
PHILIPPE CRABBE, University of Ottawa
A. L. LEVINE, University of New Brunswick

10:15 A.M. IMPACT OF GOVERNMENT RULES ON PRIVATE BEHAVIOR
Presiding: STEVEN A. Y LIN, Southern Illinois University
Papers: MICHELLE J. WHITE, New York University
Public Policy toward Bankruptcy
BARRY D. BAYSINGER AND LAURENCE S DAVIDSON, Indiana University
Rationalizing FTC Deceptive Practice Proceedings
SHARON G LEVIN, University of Missouri-St Louis, AND STANFORD L. LEVIN, Southern Illinois University-Edwardsville
The Effect of Corporate Control on the Performance of Large Nonfinancial Corporations
BERTRAND HORWITZ, State University of New York-Binghamton, AND RICHARD KOLODNY, University of Maryland
The Economic Effects of Involuntary Uniformity in the Financial Reporting of R & D Expenditures
Discussants: JEANNE WENDEL, Miami University, Ohio
ROBERT SMILEY, Cornell University
PETER D. LINNEMAN, Wharton School of Finance
F. H. DE B. HARRIS, College of William and Mary

10:15 A.M. PRODUCTION OF HEALTH (Joint Session with the Health Economics Research Organization)
Presiding: MICHAEL GROSSMAN, City University of New York and National Bureau of Economic Research
Papers: JACK HADLEY, The Urban Institute
Contribution of Medical Care to Geographical Variations in Adult Mortality Rates in the United States
STEVEN JACOBOWITZ AND MICHAEL GROSSMAN, City University of New York and National Bureau of Economic Research
Determinants of Variations in Infant Mortality Rates among Counties of the United States
ROBERT SHAKOTKO, Columbia University and National Bureau of Economic Research
Dynamic Aspects of Health, Intellectual Development, and Family Economic Status
Discussants: LOUIS GARRISON, Battelle Memorial Institute
FRED GOLDMAN, Columbia University and National Bureau of Economic Research
ROBERT MOFFITT, Rutgers University

Editor's Note:
*Papers from sessions marked with an asterisk will be published in the *Papers and Proceedings* issue of the *Review*.

ANNOUNCEMENTS

The ninety-second annual meeting of the American Economic Association will be held in Atlanta, Georgia, December 28–30, 1979. The Professional Placement Service will be open December 27–30.

The Professional Placement Service at the 1979 annual meetings of the Allied Social Science Associations in Atlanta will begin operation on December 27, the day before sessions begin. Applicants and employers will be able to attend more sessions with a day set aside entirely for labor market transactions. This service will be located at the Convention Placement Center in the Marriott Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27; 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30.

Because the 1980 annual meeting comes at an early date in the academic year (September 5–7), the Executive Committee of the American Economic Association has decided to provide employment services at a later time. A job market will *not* be organized for the September meeting in Denver, but the AEA will provide an organized market in December 1980 or January 1981 at a site yet to be selected.

Call for Papers for the 1980 Meetings

Members wishing to give papers or make suggestions for the program for the meetings to be held in Denver, September 5–7, 1980, are invited to send their ideas and abstracts to C. Elton Hinshaw, Secretary AEA, 1313 21st Avenue South, Nashville, TN 37212. Although most of the sessions sponsored by the American Economic Association will consist of invited papers, there will also be several sessions of noneconometric contributed papers. (The sessions of contributed papers will not be published in the *Papers and Proceedings* issue to appear May 1981.) Proposals for invited sessions should be submitted as soon as possible. To be considered for the contributed sessions, abstracts of proposed (noneconometric) papers must be received no later than January 15, 1980. Economists wishing to give papers on econometrics or economic theory may submit abstracts to the Econometric Society, which meets with the American Economic Association and annually schedules a substantial number of contributions.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1; for meetings scheduled between March and June, November 1. Please request application forms by writing directly to the ACLS (Attention: Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting. Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted. Awards will be announced approximately two months after each deadline.

Call for Papers. The Eastern Finance Association will hold its annual meeting, April 17–19, 1980. Topics will cover domestic and international banking, finance, and investments. Academics, business professionals, and government specialists are asked to send two-page abstracts of proposals on or before November 30, 1979 to Professor George C. Philippatos, Program Chairman, Pennsylvania State University, Department of Finance, 701 Business Administration Bldg, University Park, PA 16802

The Division of Science Resources Studies of the National Science Foundation announces the continuation of its analytical grants program. A limited number of grants will be awarded for studies focusing on aspects of training and employment of scientific and technical personnel, funding of scientific and technical activities, and outputs and impacts of scientific and technical activities. To insure funding in the 1980 fiscal year, proposals must be received by February 8, 1980. Further information may be obtained by requesting Program Announcement for the *Program for the Analysis of Science Resources*, NSF78-47 from Division of Science Resources Studies, National Science Foundation, 1800 G St., NW, Washington, D.C. 20550.

Call for Papers: On March 10–11, 1980, Brooklyn College, City University of New York, will hold a conference on "Inflation." Contact Professor Edward Marcus,

Department of Economics, Brooklyn College, Brooklyn, NY 11210.

The Sixth World Congress of the International Economics Association, "Human Resources, Employment, and Development," will be held at the Unidad de Congresos del Centro Medico Nacional, IMSS, Ave. Cuauhtemoc 330, in Mexico City, August 4–9, 1980. The Program Committee is under the chairmanship of Shigeto Tsuru (Japan), president of the IEA, with the collaboration of cochairman Samir Amin (Egypt), Rudolfo Becerril Straffon (Mexico), Mary Jean Bowman (USA), Michel Debeauvais (France), Harald Gerfin (FRG), Tigran S. Khachaturov (USSR), G. Kohlmey (GDR), Gautam Mathur (India), Franco Modigliani (USA), H.M.A. Onitiri (Nigeria), Harry Oshima (USA; Philippines), Mark Perlman (USA), Paul Streeten (U K, World Bank), Lorie Tarshis (Canada), Victor E. Tokman (Argentina), and Victor L. Urquidi (Mexico).

Arrangements are being made to provide special travel and accommodation for participants, including, when desired, pre- or post-Congress tours. Participants will receive periodic information on the progress of its preparation on request to the Local Organizing Committee (chairman Armando Labra), Colegio Nacional de Economistas, Antonio Caso 86, Mexico 4, D.F. Mexico.

Suggested contributions to the scientific work of the Congress should be submitted to the members of the Program Committee, or sent to the Secretariat of the International Economic Association, 54 Boulevard Raspail (Bureau 428), 75270 Paris Cedex 06, France, for transmission to the Committee.

One Fulbright-Hays Research Grant, tenable at the Institute of Southeast Asian Studies, Singapore, is available for each academic year until further notice. This grant is open to all American citizens with Ph.D. qualifications, who are interested in pursuing *comparative research* on topics relating to Southeast Asia within the general area of the Social Sciences and Humanities. Preference is given to candidates with proposals involving two or more individual Southeast Asian countries (or parts thereof), or the region as a whole, and who are in a position to complete their proposed projects in the stipulated twelve month period of the grant, usually commencing July 1. Travel, living, and research allowances are provided. For further information, contact Council on International Exchange of Scholars (*CIES*), 11 Dupont Circle, Washington, D.C. 20036.

The *Economic Forum* (formerly the *Intermountain Economic Review*) is proud to announce the second annual Rasmussen Prize Essay in Political Economy. The winner of this award will receive $500 and publication in the Summer 1980 issue of the *Economic Forum*. Competition for the Rasmussen Prize is open to all students who submit policy-oriented manuscripts on any area of economics. The winner of the first award is Kim-Elaine Johnson, University of Massachusetts; her article and announcement of the competition appear in the Summer 1979 issue of the *Economic Forum*. For additional information, contact Editor, *Economic Forum*, University of Utah, Salt Lake City, Utah 84112.

The John M. Olin Fellowship Program in Law and Economics, administered by the Law and Economics Center of the University of Miami School of Law, is accepting applications for the class entering in September 1980. The three-year program leads to the Juris Doctor degree, and is designed for highly motivated individuals with a strong foundation in graduate microeconomics and its application. An increasing proportion of successful candidates have the Ph.D. and have been teaching. Individuals who will have passed the preliminary examinations for the Ph.D. by September 1980 are also encouraged to apply—and to complete their dissertation while in residence at the University of Miami. Five Fellowships are available each year. These provide full tuition and fees (currently about $4,000 per year) plus stipends of $8,000 per year ($7,000 per year for Fellows without the Ph.D.). Up to five additional Fellowships of lesser amounts may also be offered each year. Deadline for submitting applications is February 15, 1980. Awards are announced by April 1, 1980. Candidates must apply separately to the School of Law and should take the Law School Admissions Test no later than December 1979. For more information write to John M. Olin Fellowship Program, Law and Economics Center, University of Miami School of Law, P.O. Box 248000, Coral Gables, Florida 33124.

The National Humanities Center will admit approximately forty Fellows for the academic year 1980–81. The Center is designed to facilitate individual research and intellectual exchange within a community of scholars. Fellows enjoy the use of private studies, conference rooms, and a dining area in the Center, and they are provided with library service and manuscript typing. The Center is located between the campuses of Duke University, North Carolina State University, and the University of North Carolina at Chapel Hill. The group of Fellows chosen each year includes both scholars of established reputation and young scholars of promise who have held the doctorate no more than ten years. In addition to scholars in fields traditionally identified with the humanities, the Center admits representatives of the natural and social sciences, the professions, and public affairs. Scholars from outside the United States are also welcome to apply. The usual term of a fellowship is the academic year, September through May, but some fellowships are available for shorter periods. Stipends are based on the normal academic salaries of Fellows; those who have partial support in the form of sabbatical salaries or research grants receive from the Center the difference between that support and their normal salaries. Fellows also receive travel expenses to and from the Center for

themselves and their families The deadline for 1980–81 applications is January 10, 1980. For application material and information write to National Humanities Center, P.O. Box 12256, Research Triangle Park, NC 27709.

Call for Papers: The Eighth Annual Telecommunications Policy Research Conference (scheduled for Spring 1980) will provide a forum for analysis and discussion of important telecommunications policy issues. Participants will include researchers and policymakers from academia, government, and industry Those engaged in research that has implications for telecommunications policy are invited to submit abstracts of 500 words or less by December 1, 1979 Authors of papers selected for presentation at the conference will be reimbursed for travel and conference living expenses if no alternative source of funding is available. Please send abstracts to: TPRC Organizing Committee c/o Robert Dansby, A.T.&T., 195 Broadway, Rm 1942B, New York, NY 10007.

The P.W.S. Andrews Memorial Prize is awarded annually for an essay by a young scholar (under the age of 30 or within 8 years of taking his first degree) in the general field of Industrial Economics and the Theory of the Firm, broadly interpreted The prize is £250 (or the equivalent in other currency) and the winning essay will normally be published in *The Journal of Industrial Economics* An essay submitted should be a work of original research by the candidate only, not previously published, and not previously awarded any other prize It should be submitted in English and should not normally exceed 10,000 words in length The closing date for entries is December 31 in each year Intending candidates for the prize should obtain details of the conditions of entry from the Administrative Officer, Office for Student and College Affairs, University House, University of Lancaster, Lancaster, LA1 4YW, England

---

Call for Papers: The Department of Philosophy of Georgia State University will hold an Interdisciplinary Conference on Capital Punishment in Atlanta, April 18–19, 1980. Academic fields represented will include philosophy, religion, sociology, criminology, law, psychology, political science, and economics. The deadline for submitted papers is December 15, 1979. Selections will be announced by February 1, 1980. Please send papers and inquiries to Professor C. G. Luckhardt, Department of Philosophy, Georgia State University, University Plaza, Atlanta, Georgia 30303.

Death

Stanley J. PoKempner, coordinator, management science programs, The Conference Board, New York, and chairman, American Marketing Association Census Advisory Committee, Nov. 13, 1978.

Visiting Foreign Scholar

Stephen J. Turnovsky: Australian National University: visiting professor of economics, University of Minnesota, fall 1979.

Promotions

Herbert J. Eskowitz: instructor of economics, Northeastern University, Sept. 1979.

Frank Falero: professor of economics and finance, California State College-Bakersfield, Mar. 15, 1979

Kenneth D. Garbade: research officer and senior economist, Federal Reserve Bank of New York, Feb. 16, 1979.

Meir Statman: assistant professor, department of economics, Rutgers College, July 1, 1978

Administrative Appointments

Barry Bressler: dean of faculty, College of Staten Island, CUNY, Feb. 1979.

Robin Carey: chairperson, department of politics, economics, and philosophy, College of Staten Island, CUNY, Feb. 1979

Appointments

Neil O. Alper, University of Tennessee: assistant professor of economics, Northeastern University, fall 1979

Shaun C. Bamford adjunct instructor, department of economics, Rutgers College, July 1, 1979.

David H Feeny, McMaster University: assistant professor of economics, Northeastern University, fall 1979.

Devra L. Golbe adjunct instructor, department of economics, Rutgers College, July 1, 1979.

David S Hoelscher, U.N. Economic Commission for Latin America. International Monetary Fund, June 1 1979.

Barry Kolatch: adjunct instructor, department of economics, Rutgers College, July 1, 1979.

Patricia V. Linton: adjunct instructor, department of economics, Rutgers College, July 1, 1979.

James C. McConnon: adjunct instructor, department of economics, Iowa State University, Feb. 7, 1979.

Robert Nakosteen, Tennessee Valley Authority. assistant professor, department of economics, James Madison University, Sept. 1979.

Tsunemasa Shiba. adjunct instructor, department of economics, Rutgers College, July 1, 1979.

V. Russell Smith, Pennsylvania State University: assistant professor, department of economics, James Madison University, Sept. 1979.

Gregory H. Wassall, University of Hartford: assistant professor of economics, Northeastern University, fall 1979.

Leaves for Special Appointment

Andrew I. Kohen, James Madison University: associate economist, Minimum Wage Study Commission, Washington, 1979–80.

Bruce M. Owen, Duke University: director, Economic Policy Office, Antitrust Division, U.S. Department of Justice, Sept. 1979–Aug. 1981.

Harold W. Watts, Columbia University: visiting senior fellow, Mathematica Policy Research, Inc., 1979–80.

Resignation

Doyle V. Peterson, Iowa State University: RFD Agri-Consultants, Feb. 28, 1979.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A Please use the following categories.

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B Please give the name of the individual (SMITH, Jane W.), her present place of employment or enrollment. her new title (if any), and the date at which the change will occur

C Type each item on a separate 3 x 5 card and please do not send public relations releases

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15; *December*, July 15.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912

Short-Run Price Effects of the Corporate Income Tax and Implications for International Trade

By JAMES R. MELVIN*

In the last decade there has been a growing awareness of the role that domestic taxation can play in determining international commodity flows and the importance of the method of taxation for the balance of payments. There has been considerable discussion on the pros and cons of value-added taxation, and of the desirability of replacing the corporate income tax with a value-added tax.[1] There has, however, been a surprising lack of empirical research,[2] and it is to the task of furthering our knowledge in this area that the present paper is addressed.

The analysis begins with a few brief comments on some theoretical issues. I first consider the question of the effect of the corporate income tax on commodity prices, for an understanding of this issue is crucial to the entire discussion. I then go on to describe how these price effects could influence trade patterns. Finally, I present the model which will be used to estimate the price effects of the tax.

I turn next to the actual calculation of price effects for the *U.S.* economy, and to carry out international comparisons, do the same calculations for Canada. I then test to see whether the calculated price effects would be expected to remain stable over the business cycle. Some comments on the importance of these results for international trade patterns are then made. It will be argued that the price effects of the corporate income tax may to a considerable extent offset the price advantages in trade that capital-rich countries are expected to have.

I. The Effect of the Corporate Income Tax on Prices and Trade Patterns

The first purpose of the analysis is to determine the relative price effects of the corporate income tax and to do this we must first ascertain how commodity prices would be expected to respond to corporate tax changes. The assumption will be that, in the short run, the tax does not affect the nominal return to the primary factors of production and that the tax is fully reflected in commodity prices. This assumption is consistent with the empirical work of Marian Krzyzaniak and Richard Musgrave.[3]

There is another branch of the literature, pioneered by Harberger which is often thought to be in opposition to the Krzyzaniak-Musgrave results, and although the model that will be used here is not completely consistent with the Harberger analysis, some comment on the relationship of these two branches of the literature seems appropriate, particularly since it has become popular to consider a variety of shifting hypotheses; an approach not employed here.[4] First, it should be noted that the shifting controversy is not about whether the tax will be reflected in commodity prices, but rather how the return to primary factors will be affected by the tax. As Harberger himself has pointed out " . . . it

*Professor of economics, University of Western Ontario This paper was extracted from a study done for the U.S. Department of Labor and the U.S. Department of the Treasury. I gratefully acknowledge their support while at the same time absolving them of any responsibility for the contents. I would like to thank John Whalley for helpful comments, Richard Hinchcliff for help in data collection, and J. Barry Smith and Francois Bourguignon for assisting with the calculations. I am also indebted to Statistics Canada for making the Canadian data available, and I am particularly grateful for the assistance of S. Nijhowne and Tony Bowen.

[1]For a review of the literature on the value-added tax, see Richard Lindholm.

[2]A notable exception is the paper by Henry Aaron, about which I will have more to say presently.

[3]For a criticism of the Krzyzaniak-Musgrave paper see John Cragg, Arnold Harberger, and Peter Mieszkowski.

[4]This, for example, is the approach taken by Aaron.

is inevitable that in the long run the corporate tax will be included in the price of the product" (p. 217). And of course as long as the capital intensities of industries differ, an increase in the price of a unit of capital must affect relative commodity prices regardless of how the tax is shifted. It is the relative price changes with which we are concerned.[5]

Given the assumption that the corporate income tax will increase commodity prices our initial empirical task will be to determine by how much. Before doing so, however, I will make some brief comments on how these price changes will be expected to affect trade patterns. Of course it is clear that for those commodities whose relative prices increase, we would expect exports to fall and imports to rise, with the opposite results for price decreases. There may well be more to it than this, however, for the corporate income tax could influence basic trade patterns as well.

The international trade literature of the last forty years has been dominated by the so-called Heckscher-Ohlin model, in which trade flows are attributed to differences in factor endowments. Although there can be no doubt that endowment differences are an important determinant of trade patterns, the emphasis on this one determinant has resulted in the neglect of other factors which may be equally important. It is clear, after all, that anything which results in differences in commodity prices between countries can give rise to trade, and it is obvious that all of the international commodity price differences one observes are not due entirely to endowment differences. One would expect such things as different production functions, differences in demand, externalities, monopolistic influences, and, of course, differences in tax structures to have an influence in determining relative world prices. Indeed it is easy to show geometrically for a fixed-coefficient model of the kind employed here, that the imposition of a tax which changes commodity prices could reverse the pattern of trade, resulting in a trade pattern opposite that predicted by the Heckscher-Ohlin model.[6]

Furthermore, if one is concerned not so much with the present structure of international commodity flows, but rather with those things that seem most likely to change that structure and thus change trade patterns, then one could argue that taxes would be the most important variable to consider. Certainly if one were interested in ways of changing trade patterns, taxes would seem to be the only viable policy tool.[7] Changes in endowments of sufficient magnitude to change trade patterns significantly would be a very long-run proposition, and it is not clear that tastes, externalities, or technology could ever be regarded as policy variables as far as trade patterns are concerned. In any case one would never seriously suggest that these variables, of such crucial importance for the entire economy, should ever be used to adjust trade flows. Such action would clearly countervene the accepted rules of matching policy goals and policy instruments. For that matter, it would not seriously be suggested that domestic taxes be imposed or reduced simply because of their influence on the trade sector There may, however, be situations in which, from a purely domestic point of view, two taxes are seen to be equivalent. In such circumstances it is clearly important that we have an understanding of the effects of both such taxes on the international sector. And, of course, even if a tax is slightly less efficient from a domestic point of view, if this deficiency is more than counterbalanced by advantages in the international sector then such a tax should be chosen. We must, after all, be concerned with overall efficiency and not just efficiency at the domestic level.

II. The Model

Let us now turn to the algebraic formulation of the model. Writing the conditions that

[5]For a full discussion of the incidence question see Mieszkowski. For a discussion of the Harberger model see my forthcoming paper. Of course I am assuming that after all shifts have occurred, the price of capital inclusive of the tax is higher than its price prior to the imposition of the tax.

[6]For a more complete discussion of the effect of commodity taxes on trade see my 1970 paper, and for a discussion specifically concerned with the corporate income tax and trade see my forthcoming paper.

[7]Note that while tariffs can change the volume of trade they cannot change the direction of trade, i.e., cannot change trade patterns.

total cost equal total revenue we have, for the jth industry,

$$(1)\quad X_jP_j = X_{1j}P_1 + X_{2j}P_2 + \ldots + X_{nj}P_n + L_jW + K_jR$$

Here P_j is the price of commodities j, and W and R are the wage rate and the return to capital, respectively. Of course this relationship holds for any production function; the special feature of this model is that a fixed relationship between the use of each of the inputs and the levels of output is assumed. Because of constant returns to scale, we can divide the jth equations by X_jP_j to obtain

$$(2)\quad 1 = a_{1j} + a_{2j} + \ldots + a_{nj} + u_j + v_j$$

Here a_{ij} is equal to $X_{ij}P_i/X_jP_j$, u_j is equal to L_jW/X_jP_j, and v_j is equal to K_jR/X_jP_j. Equation (2) shows the allocation of cost among all inputs for each dollar's worth of output.

Now suppose that a tax of rate t is applied to the returns to capital in industry j. The immediate and direct effect of this would be to raise cost and therefore unit price, and thus we would have

$$(3)\quad 1 + p'_j = a_{1j} + a_{2j} + \ldots + a_{nj} + u_j + v_j(1+t)$$

Here the per unit cost of capital has been increased by the amount v_jt and since average cost must equal average revenue, per unit price must rise by the amount p'_j. Subtracting equation (2) from equation (3) we obtain

$$(4)\quad p'_j = v_jt$$

Thus the direct effect of the tax on price is equal to the tax rate times the proportion of total cost attributable to capital.

If our model did not contain intermediate inputs, then for industry j this would be the end of the story. But with intermediate goods the tax on the capital used in industry j not only has a direct effect on the price of X_j but also has an indirect effect on the price of all other commodities. Including the indirect effect on the price of all intermediate inputs yields

$$(5)\quad 1 + p_j = a_{1j}(1 + p_1) + a_{2j}(1 + p_2) + \ldots + a_{nj}(1 + p_n) + u_j + v_j(1 + t)$$

In equation (5) the p's are interpreted as the total price changes for each of the industries. Then from equations (2) and (5) we obtain

$$(6)\quad p_j = a_{1j}p_1 + a_{2j}p_2 + \ldots + a_{nj}p_n + v_jt$$

Of course, there are similar equations for the other industries, and writing the total system in vector notation we have

$$(7)\quad p = Ap + vt$$

Here p is the vector of price changes, A is the $n \times n$ coefficient matrix, v is the vector of value-added by capital, and t is the tax rate. Solving (7) gives

$$(8)\quad p = [1 - A]^{-1}vt$$

Since $[1 - A]^{-1}$ is the inverse matrix from input-output analysis, if we know v and the tax rate t, then equation (8) gives us a simple way of calculating the price effect of the corporate income tax. It is to the task of calculating this vector of price changes that we now turn our attention.

III. Price Effects for Canada and the United States

Throughout the analysis it will be assumed that a change in the corporate income tax has no effect on the rate of return to capital or on the wage rate. As already noted this assumption is consistent with the Krzyzaniak-Musgrave model, but at variance with the Harberger analysis, for he assumes that capital will bear the burden of the corporate income tax. I have also assumed, unlike Harberger, that the corporate income tax applies to all sectors of the economy.[8]

While equation (8) seems to give a straightforward way of estimating the vector of price changes, this estimation is not without its difficulties. While the inverse matrix is readily available, the vector of value-added by capital is not. In the input-output table the capital coefficients are residuals and contain more than just the return to capital. For example, in those industries where rents exist, these rents are included. Nor is the tax to be

[8]This does not really distinguish our model from Harberger's, for our model allows the tax rate to differ between industries, and this is all that really matters.

used easy to derive, for while there is a standard rate of corporate income tax, because of such things as depletion allowances, the Domestic International Sales Corporation legislation, and the fact that many producers are not in the corporate form, the effective rate of taxation varies widely among industries.

There is one final difficulty associated with using the capital coefficients from the input-output table for the estimation of equation (8). While in the analysis I have implicitly assumed that capital is homogeneous and that the tax applies to all capital, in fact, most companies use both debt and equity capital, and the corporate income tax applies only to the returns on equity capital. Thus, for industries with a high proportion of debt capital, the capital value-added figure will result in a very poor estimate of the overall commodity price change.

Fortunately, there is a way in which all these difficulties can be overcome. The individual components of the vector vt that we seek are just the direct price effects associated with the corporate income tax. Thus, what we want are the direct price changes of the kind shown in equation (6). But given our assumption that the corporate income tax is entirely reflected in commodity prices, this vector of direct price changes is easily found by observing that the total cost of each industry has risen by the amount of tax collected, and thus the average price increase will just be total taxes divided by total sales. Thus, the vector vt can be found directly by taking total industry corporate taxes and dividing them by industry sales.[9]

As mentioned earlier, Aaron did an analysis of the effect on commodity prices of replacing the corporate income tax with a value-added tax, and before proceeding with my own calculations it will be worthwhile commenting on the relationship between Aaron's paper and the present analysis. The distinction is that his concern was with the domestic price effect of replacing the corporate tax with a value-added tax and as Aaron said, only "... passing reference will be made to the effects of such a change on United States export and import prices" (p. 162). One of the major concerns of the present analysis is the possible effects of the removal of the corporate income tax on trade. Furthermore, one of the questions with which Aaron was concerned was the sensitivity of his results to changes in the shifting assumption. I have argued that relative commodity prices will be affected regardless of the shifting assumption one makes, and thus shifting will not occupy our attention further. Finally, my analysis goes somewhat further than Aaron's in a number of respects. It uses a broader classification of industries and presents an international comparison of price changes.

Table 1 presents the direct and total effect of the removal of the corporate income tax on commodity prices for 1969 for Canada and the United States. Because the input-output tables for Canada and the United States have different aggregations, it is somewhat difficult to compare the price effects directly. I have, therefore, prepared a common aggregation. For the United States the 82-sector 1965 input-output table was used,[10] while for Canada the 1966 110-sector table was employed.[11] First, the direct effect vector was calculated by taking total corporate taxes collected for each industry as a percentage of total sales of that industry. The total effects were then calculated by use of equation (8).[12]

To derive Table 1 the Canadian and *U.S.* industries were matched as closely as possible and this produced a classification with 55 sectors for both countries. In order to produce the most accurate results possible, the aggregation was done by summing the appropriate

[9]This is not exactly correct, for this is a partial equilibrium calculation and there will be general equilibrium effects. Thus, for example, a tax on capital in industry X_j will increase the costs of all other industries using X_j as an input. This in turn will increase the price of the inputs which X_j requires. This second-order effect will be small, but nevertheless our estimates of vt will have a slight upward bias. I am indebted to John Whalley for bringing this point to my attention.

[10]The *U.S.* input-output table used is from U.S. Office of Business Economics.

[11]I am indebted to Statistics Canada for making this yet unpublished table available.

[12]The complete tables for Canada and the United States and the matching used to derive Table 1 can be obtained from the author on request.

TABLE 1—ESTIMATED PRICE REDUCTIONS ASSOCIATED WITH THE REMOVAL OF THE CORPORATE INCOME TAX FOR CANADA AND THE UNITED STATES

Industry	United States		Canada	
	Direct	Total	Direct	Total
1. Agriculture	.21	2.53	.09	1.75
2 Forestry	44	2.53	35	2.37
3. Iron Mines	6.20	8.30	1.21	4.12
4. Base Metal Mines	6.18	9.47	2.55	3.75
5 Coal Mines	.77	2.84	.11	1.77
6. Petroleum, Gas	7.02	8 81	1.96	3.59
7. Other Non-Metal Mines	2.67	5.07	3 40	5.18
8. Construction	.73	3.56	.56	3.00
9. Slaughtering and Meat Processing	2.01	4.87	2.14	4.64
10 Tobacco Manufacturers	8.17	12.09	3.75	6.66
11. Fabrics, Yarn and Thread Mills	3.29	7 99	2.47	5.43
12. Misc Textiles and Floor Coverings	69	5.03	1.66	5.19
13 Hosiery Mills	1.57	5.68	.96	3.98
14 Other Textile	1.57	7.72	1.69	4.97
15. Sawmills	2.33	4.98	2.21	4.56
16 Other Wood	2.32	5 45	1.65	4.25
17 Household Furniture	2 70	6 15	.89	3.58
18 Other Furniture	2 51	5.72	.98	3.69
19. Pulp and Paper Mills	2 74	5 87	2.60	5.02
20 Paper Box and Bag	2 73	6 31	1.42	4.79
21 Printing and Publishing	3 82	6 94	3.18	5.51
22 Chemicals and Chemical Products	4 21	8.48	3 10	5.95
23. Plastics and Synthetic Materials	4 20	8 67	1.90	5.33
24. Pharmaceuticals and Medicines	9 31	13 33	4 91	8 34
25. Paint and Allied Products	3 06	7 74	1.65	5.36
26 Petroleum and Coal Products	7 07	13 13	1 83	5.14
27. Rubber Footwear	2 99	6 59	2 97	5.68
28 Leather Tanneries	2.22	4.67	1.20	3.83
29 Shoe Factories	2 21	5.26	1 16	3 93
30. Glass and Glass Products	5 58	8.28	1.20	3.39
31 Stone and Clay Products	2 35	7.95	2.58	5.14
32 Primary Iron and Steel	1 53	4 35	2.43	5 05
33 Primary Nonferrous Metal	2 29	6.23	2 60	6.97
34. Fabricated Struc. Metal	1 98	5 33	1 70	4.08
35. Metal Stamping, Pressing	2.95	6.02	2 72	5 97
36. Other Fabricated Metal	3.54	6.72	1.96	4.89
37. Machinery and Equipment	3.26	6 61	3.92	7.04
38. Office, Computing & Acct. Machines	7 67	11.81	4.98	7.61
39. Elec Ind. Equipment	4 65	7 87	1 33	3 93
40 Electrical Appliances	3.82	8 10	.76	4.24
41. Communication Equip., Wire	2 63	5.80	1.94	5.07
42. Other Elec. Products	3.73	7.05	1.88	5.11
43 Motor Vehicles and Equip	5 10	10 69	2.46	6.55
44. Aircraft and Parts	1 89	4.76	.52	3.03
45. Other Trans. Equip	1 89	5.41	2.19	5.08
46. Misc. Manufacturing	4.06	7.13	2.74	5.52
47. Transport and Storage	1.28	3.10	1.76	3.20
48 Communications	8.56	9.48	5.84	6.85
49 Utilities	5 90	9.08	4.14	4.71
50. Wholesale and Retail	76	2.27	.67	2.19
51. Finance and Insurance	4 11	6 58	4.03	5.00
52. Hotels and Restaurants	1 02	3.19	1.36	2.86
53 Business Services	1 75	5.31	1.36	2.27
54. Auto Repairs and Services	1.21	3.53	1.36	3.08
55. Health and Educational Services	18	1.97	1 36	2.78
Arithmetic Average	3.28	6.46	2.23	4.89
Weighted Average	2.16	4.70	1.59	3.62

rows of the inverse matrix for the industries being combined so that, in effect, I have created a new input-output table for both countries. In calculating the tax variable to be applied to each of these new industry aggregations, I averaged the initial taxes paid weighted by sales. The last two rows of the table present, respectively, an arithmetic average of 3.28 and a weighted average of 2.16.[13] Eleven industries, or 20 percent, have direct effects which are larger than 5 percent.[14] It is interesting to note that these 11 include industry 3, Iron Mining, industry 4, Base Metal Mining, and industry 6, Petroleum and Gas. These have generally been regarded as industries which pay low corporation income taxes. The high ranking here, however, does not mean that the corporate tax as a percentage of *capital* is high, for these industries would be expected to be very capital intensive. Nevertheless, the fact that the corporate income tax results in relatively large price increases for the outputs of these industries is of interest, for these are important intermediate products.

It is also noteworthy that these top 16 industries include industry 38, Office, Computing and Accounting Machines. There is some tendency for the very highly sophisticated manufacturing sector to incur relatively large direct price effects.[15] These are industries which, for technically advanced countries such as the United States, would be expected to generate exports, and it would appear that the corporate income tax may be offsetting the comparative advantage which we would expect to observe. Finally, note that industry 43, Motor Vehicles and Equipment, is in this group of the top 11 industries.

Turning to the total price effect of the corporate income tax, we observe a significant increase in prices due to the use of intermediate products. We find that the arithmetic average has increased from 3.28 to 6.46 while the weighted average has increased from 2.16 to 4.70. Also of interest is the variation which exists between the direct effect and the total effect on an industry-by-industry basis. Some of the industries with relatively high direct effects, such as industry 38, Office, Computing and Accounting Machines, and industry 48, Communications, show rather modest increases; in both cases less than 40 percent. On the other hand, some of the sectors where the direct effects are small show relatively large increases, and in fact in industries 1 and 55, the increase is over 1,000 percent. Of particular interest is the fact that industry 1, the basic agricultural sector, while paying very low corporate income taxes, nevertheless has a total price increase of approximately 2.5 percent due to the corporation income tax structure.

It is also of interest to note that some of the total price effects are quite high in absolute terms. There are 5 industries, or approximately 10 percent of the total, in which the total price effect is over 10 percent. These, in general, tend to be the same industries that had high direct effects. It is noteworthy that for industry 43, Motor Vehicles and Equipment, the price increase associated with the corporate income tax is 10.7 percent. It seems obvious that a decrease of over 10 percent in the price of *U.S.* produced automobiles would substantially improve the competitive position of the automobile industry in world markets.

The third and fourth columns of Table 1 present for the Canadian economy the same calculations as were reported for the United States in the first two columns. The numbers for Canada tend to be lower than for the United States; the arithmetic average for the total effect for Canada is 4.9, as compared to 6.5 for the United States, and the weighted

[13]These averages have been calculated from the original input-output tables for both countries and not for the aggregated table. Note that in the calculation of these averages the dummy industries of the input-output tables, industries 78, 79, 80, and 81 in the *U.S.* table and industries 108, 109, and 110 of the Canadian table, have been excluded.

[14]For the full *U.S.* table the range is even larger with industry 51, Office, Computing and Accounting Machines having a direct effect of 12 percent. In the aggregation *U.S.* industries 51 and 52 were combined resulting in the smaller figure shown for industry 38 of Table 1.

[15]Other sophisticated manufacturing industries which are in the top 20 percent in the full *U.S.* table are Scientific Instruments and Optical Equipment. In Table 1 these are included in industry 46, Miscellaneous Manufacturing.

averages are 3.6 for Canada and 4.7 for the United States. In the United States there were 11 industries, or approximately 20 percent, where the direct effects were in excess of 5 percent; for the Canadian table there is only one such industry.[16] Turning to the total effects we find no industries in which the total price increase is more than 10 percent, as compared to the 5 industries in the United States.[17]

Considering specific industries we note that the direct price effects for the United States are almost always higher than they are for Canada and in some cases the differences are quite significant. For industries 3 and 4, the mining sector, we see that the price changes in the United States are, on average, more than twice what they are in Canada. And for industry 6, the petroleum and natural gas producing sector, the price reductions in the United States are 8.8 percent as compared to 3.6 percent in Canada. The industries 23–26 show much the same kind of relative price differentials. The removal of the corporate income tax would make a modest contribution to the effort currently being waged to reduce petroleum prices. Also note that for industry 38, Computers and Office Machines, and for 43, Motor Vehicles, the price reductions are significantly higher in the United States, suggesting that a beneficial stimulus to the exports of these industries by the United States could occur if the corporate income tax were removed.

The significant difference which exists between the Canadian and *U.S.* price effects is interesting, and although the calculations presented here cannot cast much light on why there will be such differences, a number of possible explanations come to mind. Of course higher capital intensities in *U.S.* industries would produce this result, but while it would generally be agreed that the United States is more capital intensive than Canada, it seems unlikely that this difference could cause such significant differences. Certainly for industries such as Motor Vehicles and Equipment, which are known to be quite similar between Canada and the United States, some further explanation is required.

Another factor which may in part explain the price differences would be differences in the debt-equity ratios between the two countries. It is generally believed that Canadians are more risk averse than their *U.S.* counterparts, and if this is true, Canadians may demand a higher risk premium for equities. This would make debt financing relatively cheaper in Canada than in the United States and would result in a higher debt-equity ratio in Canada. It also seems likely that the large amount of foreign ownership of Canadian industry contributes to higher debt-equity ratio. Foreign parent firms may encourage the use of debt capital in a higher proportion than is the case for Canadian owned firms; it is well known that certain *U.S.* parents have historically been reluctant to raise equity capital in Canada. It is noteworthy that in many of the industries where there is substantial ownership of Canadian companies by *U.S.* parents, the differences between the direct price effects are quite significant. As examples, see industries 3, 4, 6, 23, 24, 25, 26, 38, and 43.

One question which comes to mind when considering the total price effect attributable to the corporate income tax is the sensitivity of these figures to the business cycle. In particular, we are interested in knowing whether or not these figures vary significantly as aggregate corporate profits change. If there is a wide variability from year to year, then the particular year chosen could lead to misleading conclusions if this tends to be a particularly high or low corporate profit year. Furthermore, if there is considerable variability in the total price effect of the corporate income tax, then the removal of this tax would be expected to result in larger price fluctuations over the business cycle. In order to investigate the sensitivity of the total price effects to the business cycle, calculations similar to those shown in Table 1 were done for three years, 1965, 1969, and 1971, and the averages are presented in Table 2.[18] The

[16]For the full Canadian table there are 4 such industries, or approximately 4 percent of the total.

[17]In the full Canadian table, only Distilleries with a total price effect of 17.55 is over 10 percent.

[18]The full tables for these and other reported calculations are available from the author on request.

TABLE 2—AVERAGE PRICE REDUCTIONS ASSOCIATED WITH THE REMOVAL OF THE CORPORATE INCOME TAX FOR CANADA

	1965		1969		1971	
	Direct	Total	Direct	Total	Direct	Total
Arithmetic Average	2.90	6.55	2.90	6.37	2.60	5.71
Weighted Average	1.93	4.63	2.10	4.74	2.03	4.51

Canadian numbers were used because the data on corporation income tax payments at the required industry level were not available for the United States for any year other than 1969. These three years were chosen because 1965 represents a year in which corporate profits were high, while 1971 was a year of low corporate profits, and of course 1969 was chosen because this was the year for which the *U.S.* calculations were made.

The remarkable aspect of the results shown in Table 2 is the very small variation in the total price effect which exists for these three years. And not only are the averages almost the same, there was also surprisingly little variability among the industries. It should be noted that the average numbers for 1969 for Canada presented in Table 2 are not comparable with the numbers in Table 1 for in Table 2 we are using total federal and provincial corporate income taxes collected as opposed to only federal taxes in Table 1. Unfortunately, federal taxes alone were not available in Canada for 1971. There does not, however, seem to be any reason to expect that the variability would be any greater for federal taxes alone than it is for the total corporate taxes. Of course these results for Canada do not necessarily imply that the same conclusions hold for the United States. There does not seem to be any reason, however, to expect that the situation for the United States would be substantially different. It thus seems fair to conclude that the price effects of the corporate income tax remain quite stable over the business cycle.

The purpose of presenting the calculations for both Canada and the United States was to get some estimate of how the price reductions calculated for the United States would compare to price changes which would occur in other countries if the corporate tax were removed there as well. I have found that even for Canada, which is likely to be among the most capital-intensive trading partners of the United States, and where the corporate income tax rates are very similar, there are significant differences in the price effects.[19] For those countries which are both less capital intensive and which have lower rates of corporate income tax, the difference would be expected to be larger. And it must be remembered that higher domestic prices not only disadvantage exporting firms, they also cause hardship for import-competing industries. As one example, consider the textile and clothing sector, industries 11–14. Although the price effects here are not extreme, they are above average, being 6.6 unweighted and 6.2 weighted. One suspects that foreign exporters of these products would have low corporate tax price effects, perhaps close to zero, and thus the *U.S.* tax is equivalent to a 6 or 7 percent subsidy to foreign producers. The removal of the corporate income tax, then, would be equivalent to a tariff increase of 6 or 7 percent for this sector.

The most popular explanation of trade patterns in the last few decades has been the Heckscher-Ohlin model, which attributes trade to differences in endowments. The essential argument of the theorem is that when a factor of production is plentiful and therefore relatively inexpensive, those products which use relatively large amounts of this

[19]It is difficult to get an exact comparison of the rates, for while the quoted rates are almost the same, both countries allow deductions for state or provincial corporate tax payments. For Canada the provincial corporate tax which is deductible from federal taxes ranges from 10 to 15 percent, but in the United States the allowable deductions vary. It would, of course, have been preferable to use total taxes collected, but such data were not available for the United States.

factor will also tend to be relatively less expensive, and we would thus expect the export bundle to be composed predominantly of such commodities. The United States is regarded as being capital intensive, and thus we would expect those commodities which have a high capital component to be relatively less expensive than in foreign countries and, *ceteris paribus*, be exported.

We have seen, however, that the corporate income tax acts to offset this comparative advantage, for the higher is the capital component of output the higher will the price effect of the corporate income tax tend to be.[20] Indeed it is quite possible that for some industries the corporate income tax may completely offset any advantage due to capital abundance. In terms of our fixed-coefficient model, suppose that production functions in two countries are the same but that the price of capital is relatively lower in the home country than in the other. Further suppose that there is a 50 percent corporate income tax in the home country and no tax in the other. If all other prices were the same, the cost of capital could be 50 percent higher in the foreign country and still generate the same output price. Clearly, a higher cost of capital will have the same effect on commodity prices regardless of whether it is due to a tax or to a higher price for capital services. Of course in practice other countries do have corporate income taxes, although they are often less than in the United States, and production functions are not identical among countries. The point remains, however, that it is quite possible that the corporate income tax could offset the advantage which might exist in international markets due to lower capital costs. It seems quite possible that the failure of investigators to show that *U.S.* exports are capital intensive relative to imports, the well-known Leontief Paradox, may well be due, at least in part, to the effects of the corporate income tax.

IV. Conclusions

The purpose of this paper has been the estimation of the commodity-price effects of the corporate income tax through the use of the input-output table. It was found that for the United States the price changes ranged from 2–16 percent, the mean being 4.7 or 6.5 depending on whether the average was weighted or unweighted. The first major conclusion is that not only does the corporate tax result in a significant price increase, but more importantly, that the price effect varies significantly among industries. These differential price effects have doubtlessly resulted in significantly different domestic production and consumption patterns than would have existed had the same amount of revenue been raised through a uniform rate of sales or value-added tax.[21] There has been, in other words, a considerable amount of resource misallocation associated with the corporate income tax. It is difficult to think of any justification for this very arbitrary form of economic discrimination.

Turning to the international trade effects, it was found that the corporate income tax tends to raise all commodity prices, and since the General Agreement on Tariff and Trade (*GATT*) regulations do not allow the kinds of rebates for corporate income taxes that are allowed for sales and excise taxes, and since imports clearly are not subject to this tax, the overall effect must be to increase imports and decrease exports. Of course such aggregate effects can, in theory, be neutralized by appropriate changes in exchange rates, but at least for the United States such changes have been anything but easy.

Perhaps the most important effect on the trade side, however, is the fact that because the amount of the price increase is associated with the capital intensity of the industry, the tax has tended to discriminate against capital-intensive products; exactly those products in which the United States would be expected to have a comparative advantage. Thus the

[20]There will not be a perfect relation between physical capital intensities and corporate tax liability, for the corporate income tax applies only to equity and not to debt capital. Furthermore, any tax concessions granted specific industries will affect this relationship.

[21]A value-added tax of approximately 5.6 percent would have, in 1969, raised the same amount of revenue as the corporate income tax.

corporate income tax acts to offset the trade patterns associated with the unequal distribution of resources in the world and acts to prevent the realization of the gains associated with international trade.

REFERENCES

H. Aaron, "The Differential Price Effects of a Value-Added Tax," *Nat. Tax J.*, June 1968, *21*, 162–75.

J. G. Cragg, A. C. Harberger, and P. Mieszkowski, "Empirical Evidence on the Incidence of the Corporation Income Tax," *J. Polit. Econ.*, Dec. 1967, *75*, 811–21.

A. C. Harberger, "The Incidence of the Corporate Income Tax," *J. Polit. Econ.*, June 1962, *70*, 215–40.

Marian Krzyzaniak and Richard A. Musgrave, *The Shifting of the Corporation Income Tax*, Baltimore 1963.

R. W. Lindholm, "The Value Added Tax: A Short Review of the Literature," *J. Econ. Lit.*, Dec. 1970, *8*, 1178–89.

R. I. McKinnon, "Intermediate Products and Differential Tariffs: A Generalization of Lerner's Symmetry Theorem," *Quart. J. Econ.*, Nov. 1966, *80*, 584–615.

James R. Melvin, "Commodity Taxation as a Determinant of Trade," *Can. J. Econ.*, Feb. 1970, *3*, 62–78.

———, "Some General Equilibrium Consequences of the Corporate Income Tax," *J. Publ. Econ.*, forthcoming.

———, *The Tax Structure and Canadian Trade: A Theoretical Analysis*, Ottawa 1975.

P. Mieszkowski, "Tax Incidence Theory: The Effect of Taxes on the Distribution of Income," *J. Econ. Lit.*, Dec. 1969, *7*, 1103–24.

R. D. Warne, "Intermediate Goods in International Trade with Variable Proportions and Two Primary Inputs," *Quart. J. Econ.*, May 1971, *85*, 225–36.

U.S. Office of Business Economics, "Input-Output Structure of the U.S. Economy: 1963," *Surv. Curr. Bus.*, Nov. 1969, *45*, 16–47.

The Migration Decision: What Role Does Job Mobility Play?

By ANN P. BARTEL*

An important characteristic of the *U.S.* population is its geographic mobility. In 1970, 18 percent of the population was living in a county that was different from their 1965 county of residence; half of these migrants had also moved across state lines.[1] Previous work on geographic mobility can be classified into two categories.[2] The first is composed of studies that have used aggregate data (for example, Samuel Bowles, Michael Greenwood, 1969, Ira Lowry, and Aba Schwartz) to examine the determinants of net or gross migration for *SMSA*s or other geographic divisions. The second category of research has used data on individuals (for example, Julie DaVanzo, Richard Kaluzny, John Lansing and Eva Mueller, and Solomon Polachek and Francis Horvath) to explore the relationship between an individual's characteristics and his decision to migrate.

This article continues the work on the analysis of the individual's decision to migrate, but differs from the previous studies by focusing on the relationship between job mobility and migration. First, the proportion of geographic mobility that occurs in conjunction with a job change is calculated. Second, it is shown that the true effects of human capital variables, job characteristics, and family variables on the decision to migrate are best measured when one takes account of the relationship between migration and job mobility. Third, the effect of migration on th wage gains of individuals is studied and agai the need for distinguishing among moves tha were associated with quits, layoffs, and trans fers is clearly shown. Finally, by using thre data sets that encompass different age group (the National Longitudinal Surveys (*NLS*) o Young and Mature Men and the Coleman Rossi Retrospective Life History Study), th importance of the relationship betwee migration and job mobility is demonstrated a different points in the life cycle.[3]

Section I of the article presents som summary statistics on the extent of geo graphic mobility among the individuals in th samples and documents the relationshi between migration and job mobility. I Section II a framework for analyzing th decision to migrate is discussed. Sections II and IV present the empirical results whil Section V summarizes the analysis.

I. Some Evidence on Migration and Job Mobilit

Table 1 contains summary statistics on th rate of migration and job mobility in the thre data sets. In the case of the two *NLS* sample migration is defined as a move to a differen Standard Metropolitan Statistical Are (*SMSA*) or county, while for the Coleman Rossi individuals migration is a move acros state lines. The Coleman-Rossi migratio rates are therefore expected to be low relativ to intercounty migration rates. Column presents data for the *NLS* Young Men for th period 1971–73. These men were between th

*Columbia University Graduate School of Business and National Bureau of Economic Research. I would like to thank Jacob Mincer and Robert Pollak for their very helpful comments on a previous draft of this paper. This research was supported through the National Bureau of Economic Research by a grant from the National Science Foundation (Grant No. SOC71-03783 A03). The opinions expressed herein are my own and do not necessarily reflect the views of the National Science Foundation.

[1]See the 1970 Census of Population, General Social and Economic Characteristics, Table 87.

[2]See Michael Greenwood (1975) for a summary of the literature on geographic mobility.

[3]The National Longitudinal Survey (*NLS*) of Matur Men is described in U.S. Department of Labor (1970 74), the *NLS* of Young Men is described in U.S Department of Labor (1970–77), while the Coleman Rossi data, which were collected in January 1969, ar discussed in Z. Blum, N. Karweit, and A. Sorensen. Th analysis is restricted to the white men in all thre samples.

TABLE 1—DESCRIPTION OF GEOGRAPHIC MOBILITY

	NLS Young Men 1971–73	Coleman-Rossi 1964–69	*NLS* Mature Men 1966–71
Proportion who moved	20	.13	.06
Proportion of moves involving quits	.49	.53	38
Proportion of moves involving layoffs	.13	14	26
Proportion of moves that did not involve a job change	38	33	.36
Proportion of quitters who migrated	32	20	16
Proportion of those laid off who migrated	26	19	07
Proportion of job stayers who migrated	12	06	03
Proportion of job changer migrants who moved for economic reasons	.76[a]	.75[a]	.52
Proportion of migrants who moved for economic reasons	.85[b]	83[b]	69[b]
Proportion of moves "caused by" the decision to change jobs	47[c]	50[c]	.33[c]
Sample Size	1,608	580	1,790

[a]Based on data in Lansing and Mueller for similar age groups.
[b]Includes transfers
[c]See text for method used to calculate this statistic

ages of 19 and 29 in 1971. In column 2 data are shown for the 1964–69 period for the Coleman-Rossi individuals who were between the ages of 26 and 35 at the start of the period. Column 3 contains data for the *NLS* Mature Men for the period 1966–71. These men were between the ages of 45 and 59 at the start of the period. A shorter time interval was used for the *NLS* Young Men in order to maximize the number of individuals who were not enrolled in school.[4]

The second, third, and fourth rows in Table 1 document the relationship between migration and job mobility. For all samples we observe that roughly two-thirds of all moves involved a job separation.[5] To determine whether these moves were in fact caused by the decision to change jobs, information on the reasons for migration is necessary. The *NLS* Mature Men data set provides information on whether a move was undertaken for economic reasons or personal (for example, family, health) reasons. One can argue that moves that involved job separations and that were made for economic reasons were in fact caused by the decision to change jobs. Those moves that involved job separations but were made for personal reasons can be said to have caused the accompanying job separations. For the *NLS* Mature Men, 52 percent of those individuals who migrated and separated indicated that they moved because of economic reasons. In order to estimate what percentage of *all* moves are *caused* by the decision to *change* jobs, however, the number of individuals who separated *and* who migrated for economic reasons as a percentage of *all* migrants is calculated. For the mature men, this proportion is one-third. For the younger samples, no information is provided on the reasons for migration. Data from another source, however, enables us to make similar calculations for these age groups.[6] As Table 1 indicates, 75 percent of those individuals who migrated and separated moved because of economic reasons resulting in one-half of all moves in the younger cohorts being *caused* by the decision to change jobs. This analysis, therefore, indicates the importance of studying the decision to migrate in conjunction with the decision to separate from a firm.

[4]For example, in 1967, 50 percent of the sample was enrolled in school, but by 1971, only 17 percent of the sample was enrolled. At the time this study was done, data for the *NLS* Young Men were available only up to 1973.

[5]Note that for the two younger samples, 80 percent of these separation-related moves are due to quits, while for the older sample 60 percent are due to quits. These differences across samples are, of course, related to the decline in the ratio of quit rates to layoff rates with age

[6]Lansing and Mueller find that 77 percent of individuals aged 18–24 who migrated and separated during a five-year period moved because of economic reasons while the same statistic is 75 percent for men aged 25–34 See their Table 9.

II. Theoretical Framework and Empirical Specifications

The data presented in Section I document the relationship between migration and job mobility. In this section the framework within which the decision to migrate can be analyzed is discussed and it is shown how job separations can be integrated into this analysis.

Economic theory predicts that an individual will attempt to sell his services in the market which offers him the highest return. Larry Sjaastad utilized this basic concept in his analysis of internal migration in the United States. The individual is guided by his discounted net return from migrating at time t; if this net return is positive, he will migrate. In other words,

$$PM_t = f(G_t) \tag{1}$$

where PM_t is the probability that the individual moves in time period t and G_t is the discounted net gain from moving. Thus G_t can be written as follows:

$$G_t = Y_t^* - Y_t - C_t \tag{2}$$

where Y_t^* is the present value of the expected real income stream if the individual migrates in time period t, Y_t is the present value of the expected real income stream in the current location calculated at time t, and C_t are the costs of migration as well as such costs as the loss of the wife's earnings (assuming such a loss occurs), the costs of uprooting school-age children as well as the time costs of searching for a job and residence in the new location. If $G_t > 0$, the individual is assumed to migrate.

The probability PM_t can be viewed as the unconditional probability of migration. As was shown in Section I, some migrants are also job quitters, some were laid off, and the remainder are individuals who did not change employers. In other words, PM can be viewed as the sum of three joint probabilities:

$$PM = P(Q \cap M) + P(L \cap M) + P(NS \cap M) \tag{3}$$

where

$P(Q \cap M)$ = the joint probability of quitting and migrating

$P(L \cap M)$ = the joint probability of being laid off and migrating

$P(NS \cap M)$ = the joint probability of not separating from the firm and migrating

Moreover each joint probability can be rewritten as

$$P(X \cap M) = P(X) \cdot P(M|X) \tag{4}$$

where $X = Q$, L, or NS,[7] $P(X)$ is the probability of a quit, layoff, or no separation, and $P(M|X)$ is the probability of migration conditional on a quit, layoff, or no separation. Equation (4) shows how the decision to migrate is directly linked to the probability of a job separation. In studying migration, therefore, we see that it is crucial to have an understanding of both the process of job mobility and the determinants of the conditional probabilities of migration.

In order to study the determinants of the probability of migrating, those variables which measure the discounted net return from moving must be identified. Since the decision to move has already been shown to be closely tied to the decision to change jobs, the analysis is in part an attempt to measure the discounted net return from changing jobs. In addition, those variables which determine the conditional probability of migration must be examined. Of course, there may be some overlap in the sets of variables that determine $P(X)$ and $P(M|X)$. More important is the fact that some variables may affect the discounted net return from migration only because they affect the discounted net return from separating. This points out the importance of examining the joint probabilities of migration rather than the unconditional probability of migrating. Moreover, it suggests that a convenient way of determining whether

[7]A move that did not involve a job separation could either be an intrafirm job change or residential mobility. Although the data do not distinguish between these two types of moves, the fact that migration is defined as a move to a different *SMSA*, county, or state indicates that most of these moves are probably transfers. In the remainder of the article, this assumption is maintained.

a variable's measured effect on migration is due solely to its effect on the probability of changing jobs is to compare the effect of that variable on the probability of separating and migrating with its effect on the probability of separating and not migrating (i.e., changing jobs in the local labor market).

For example, consider the effect of the individual's current wage. According to the theory of specific training, the wage should have a negative effect on the probability of quitting, but an ambiguous effect on the probability of a layoff.[8] If the wage does not affect the conditional probability of migration, then a negative effect of the wage on the probability of migrating may be observed only if the individual quit his job. And, if the wage affects migration simply because it affects the probability of separating, then the measured effect of the wage on the joint probability of migrating and separating would be the same as its effect on the probability of separating in the local labor market. In the case of transfers, however, the wage should have a positive effect on migration since employers would be likely to transfer (i.e., invest in) those individuals who already have a large amount of specific training and are closely tied to the firm.[9] This analysis shows, therefore, that the true effect of the wage on the probability of migration can only be estimated when the different types of moves are distinguished from one another.

Other variables can be suggested as determinants of the discounted net gain from moving. Education should have a positive effect on the conditional probability of migration since more highly educated individuals would tend to have better information about nonlocal job opportunities, may be more adaptable to change, and tend to be in occupations that operate in a national labor market. This would predict a stronger effect of education on nonlocal separations than local separations. Whether education will in fact have an effect on the unconditional probability of migration, however, is unclear For example, Borjas and I present evidence that more educated individuals are significantly less likely to be laid off. This would suggest that in the case of the joint probability of being laid off and migrating, the effect of education is ambiguous in sign. Therefore in estimating the relationship between education and migration one would want to distinguish among types of moves.

One of the most important sets of determinants of the net return from migration is the characteristics of the individual's family.[10] For example, married individuals with working wives should have higher costs of migration than those whose wives are not in the labor force. Similarly, individuals with school-age children should have a lower net return from migration, everything else held constant. Again, however, the effects of these variables on the decision to migrate may depend on the association of a move with a job separation. For example, the presence of a working wife may have little effect on the probability of being transferred by one's employer; the true inhibiting effect of wife's

[8]The theory argues that employees with more worker-financed specific training are less likely to quit and those with more firm-financed specific training are less likely to be laid off. Following Donald Parsons, since an individual's wage can be expressed as

$$W = a_0 + a_1E + a_2S_w$$

where E = education and S_w = worker-financed specific training, the quit probability will be inversely related to the wage when education is held constant The sign of the relationship between the layoff probability and the wage depends on whether firm-financed specific training is positively or negatively correlated with worker-financed specific training. Although the positive correlation is more likely, the layoff probability may still be positively related to the wage if job instability is compensated by a wage premium. See Robert Hall for a discussion of the relationship between wages and separation rates according to the theory of compensating wage differentials George Borjas and I discuss additional theories that can be used to explain the wage rate-separation rate relationship.

[9]The equation in fn. 8 shows that the individual's wage is positively correlated with the amount of specific training he possesses since an individual with more worker-financed specific training is also likely to have more specific training in total. In other words, the incentives that exist for the worker to invest in specific training are also likely to induce the firm to invest in the worker

[10]See DaVanzo, Larry Long, Jacob Mincer, and Polachek and Horvath for empirical evidence.

labor force participation on the migration decision of job separators would then not be correctly estimated in an analysis that did not distinguish among types of moves.

This approach also indicates why certain job-related characteristics should affect the net return from migration. For example, individuals with low levels of tenure in the current job are more likely to experience a job separation.[11] And, given the underlying relationship between migration and job separations, these individuals would therefore be likely to move geographically. More important is the notion that the correlation between job tenure and length of residence may produce the observed negative effect of residence when in fact the true causative variable is job tenure.

The analysis presented here thus shows that the relationship between job mobility and migration can be demonstrated by estimating the following set of linear probability equations:

$$PM_t = a(Z_i, F_i, J_t) \tag{5}$$

$$P(Q \cap M)_t = b(Z_t, F_t, J_t) \tag{6}$$

$$P(L \cap M)_t = c(Z_t, F_t, J_t)$$

$$P(NS \cap M)_t = d(Z_t, F_t, J_t)$$

$$P(Q \cap NM)_t = e(Z_t, F_t, J_t) \tag{7}$$

$$P(L \cap NM)_i = f(Z_t, F_t, J_t)$$

where Z_t is a vector of individual characteristics, F_t is a vector of family characteristics, J_t is a vector of job-related characteristics, and NM means not migrating. It has been argued that the coefficients in equation (5) will not correctly estimate the effects of the independent variables for all movers since, as equations (6) show, there are three distinctly different types of moves. Further, a comparison of equations (6) and (7) will show whether an independent variable affects the probability of migrating simply because it determines the probability of separating, that is, whether it is useful to distinguish local separations from nonlocal separations.

[11]See the papers by Borjas and myself and by Boyan Jovanovic and Mincer.

III. Empirical Findings on the Determinants of Migration

In this section the results of estimating equations (5), (6), and (7) using data from the *NLS* of Young Men, the Coleman-Rossi Retrospective Life Histories Study, and the *NLS* of Mature Men are presented. Table 2 contains the regressions for the *NLS* Young Men sample while Table 3 contains similar regressions for the Coleman-Rossi sample and Table 4 presents the regressions for the *NLS* Mature Men.[12,13] The regressions in these tables do not hold job tenure and length of residence constant since it can be argued that these variables are serially correlated with the dependent variable, that is, previous moves have determined current job tenure and length of residence. In fact, when tenure and residence are added to the regressions, some of the other independent variables do become less significant (but the conclusions of this analysis are unchanged) indicating that these variables also determined previous mobility. The coefficients on tenure and residence from these regressions are shown in Table 5 and the complete regressions are available from me upon request. In order to focus on the distinction between unconditional and joint probabilities of migration, each independent variable is discussed in turn to show how its measured effect on migration depends on the associated job separation. The variables are defined in Table 2; note that the independent variables are measured at the beginning of the period under study.

A. *The Wage*

As discussed in Section II, the effect of the individual's wage rate on the probability of migrating may depend on whether migration

[12]For the *NLS* Young Men the time period under study is 1971–73, for the Coleman-Rossi sample it is 1964–69, and for the *NLS* Mature Men it is 1966–71. See fn. 4 for a discussion of the reason that the *NLS* Young Men analysis was restricted to a two-year period.

[13]Since the dependent variables are dichotomous, ordinary least squares is not the proper estimating technique. This article therefore utilizes maximum likelihood logit; the coefficients presented in the tables are the marginal

TABLE 2—DETERMINANTS OF 1971–73 MIGRATION: *NLS* YOUNG MEN[a,b]
MAXIMUM LIKELIHOOD LOGIT ESTIMATES

	Uncond	*Migqt*	*Miglay*	*Migtr*	*Nmquit*	*Nmlay*
EDUC	.0127	0057	.0023	.0044	−.0222	−.0064
	(2.37)	(1.45)	(1.08)	(1.25)	(−3.98)	(−1.79)
EXPER	−.0126	−.0102	−.0018	−.0011	−.0070	−.0003
	(−3.18)	(−3.36)	(−1.09)	(−.45)	(−1.84)	(−.11)
WAGE	.0060	−.0059	0006	.0099	−.0136	− 0049
	(.89)	(−1 10)	(22)	(2.43)	(−1.73)	(−.97)
MAR	−.0223	0155	−.0088	−.0265	−.0381	−.0263
	(−.80)	(.76)	(−.79)	(−1.42)	(−1.34)	(−1.47)
WLFP	.0526	0035	.0043	.0419	.0438	−.0248
	(1.68)	(15)	(.32)	(2.09)	(1.35)	(−1 09)
WINC	−.0009	− 0002	−.0001	−.0006	−.0002	.0003
	(−1.51)	(− 45)	(− 41)	(−1 46)	(−.36)	(.67)
SCHL	.0071	.0185	− 0138	− 0029	−.0392	.0127
	(.23)	(.78)	(−.84)	(−.15)	(−1.18)	(.61)
HLTH	.0692	.0024	−.0013	0568	−.0078	0090
	(1.98)	(.09)	(−.08)	(2.87)	(−.19)	(36)
UNEMP	−.0048	0181	0161	− 0529	0659	.0587
	(− 20)	(1.09)	(1.91)	(−2 82)	(2.97)	(4.27)
χ^2	50.50	32.41	16.19	34 57	63.78	42 37
N	1608	1608	1608	1608	1608	1608

[a]Asymptotic *t*-ratios are given in parentheses. Definition of variables are *EDUC* = years of education, *EXPER* - potential experience (as of 1971) since completion of schooling (*NLS* Young Men), *REM* - time remaining until retirement as of 1966 (*NLS* Mature Men), *AGE* - age in 1964 (Coleman-Rossi), *WAGE* = hourly (*NLS*) or monthly (Coleman-Rossi) wage at the beginning of the period, monthly wage is in tens of dollars, *MAR* - one if individual is married, *WLFP* - one if individual's wife was in the labor force at the beginning of the period under study, *WW* - wife's hourly wage (*NLS* Mature Men), *WINC* - wife's earnings in hundreds of dollars (*NLS* Young Men), *SCHL* = one if individual has school-age children, *HLTH* = one if individual's health limits kind or amount of work (*NLS* Young and Mature Men), *UNEMP* - one if individual unemployed during the previous year, *JOB* = job tenure in years at the beginning of the period and *RTEN* - the difference between length of residence and job tenure at the beginning of the period.

[b]Column headings refer to the probability of migration· *Uncond* is the unconditional probability, *Migqt* is the joint probability of migrating and quitting, *Miglay* is the joint probability of migrating and being laid off, *Migtr* is the joint probability of migrating and not changing employers; *Nmquit* is the joint probability of not migrating and quitting, *Nmlay* is the joint probability of not migrating and being laid off

is associated with a job separation. The results in the first column of each table show that the wage has no effect on the unconditional probability of migration. The reason for this somewhat paradoxical result is made clear by an examination of the other regressions in Tables 2, 3, and 4. We find that *WAGE* has a negative effect (which is significant only for the *NLS* Mature Men) on the probability of migrating and quitting in all three samples. However, in the case of the joint probabilities of migrating and being laid off or migrating and not changing employers, the wage coefficient is always nonnegative and in some cases is significant. The reason for this nonnegative wage effect was suggested in Section II. Since the joint probability of migrating is a function of the probability of separating, the nonnegative wage coefficient may be due to a nonnegative wage effect on the probability of being laid off and the probability of being "promoted" via a transfer. In fact, for the younger cohorts (both *NLS* and Coleman-Rossi), transfers depend positively and significantly (in the case of the *NLS* Young Men) on the wage level.

Is the negative effect of the wage on the probability of migrating and quitting due solely to the negative relationship between

effects of the independent variables on the dependent variable, evaluated at the mean of the dependent variable.

TABLE 3—DETERMINANTS OF 1964–69 MIGRATION: COLEMAN ROSSI[a] MAXIMUM LIKELIHOOD LOGIT ESTIMATES

	Uncond	*Migqt*	*Miglay*	*Migtr*	*Nmquit*	*Nmlay*
EDUC	.0136	.0049	.0008	.0082	−.0048	−.0073
	(2.62)	(1.25)	(.41)	(2.70)	(−.74)	(−2.07)
AGE	−.0009	−.0010	−.0002	.0002	−.0014	.0003
	(−2.04)	(−2.76)	(−1.09)	(.90)	(−2.33)	(.85)
WAGE	.0003	−.0006	.0001	.0004	−.0006	.0002
	(.46)	(−1.08)	(.79)	(1.52)	(−.71)	(.53)
MAR	−.0371	−.0021	−.0096	−.0222	−.0381	.0045
	(−.96)	(−.07)	(−.71)	(−.94)	(−.67)	(.13)
WLFP	−.0210	−.0253	−.0065	.0119	.0749	−0167
	(−.58)	(−.92)	(−.43)	(.57)	(1.68)	(−.59)
SCHL	.0010	.0083	−.0078	.0026	.0404	−.0043
	(.03)	(.31)	(−.50)	(.13)	(.90)	(−.17)
UNEMP	.1808	.0646	0452	.0294	−.1624	.0986
	(2.73)	(1.26)	(2.99)	(.68)	(−1.04)	(2.08)
χ^2	21.24	13.45	10.15	14.81	12.97	8.72
N	579	579	579	579	579	579

[a]Asymptotic *t*-ratios are given in parentheses. Variables are defined in Table 2

wages and quitting? This question can be answered by looking at the regressions on the probability of quitting and not migrating. The results show that the wage effect in these equations is as strong or stronger than the effect in the associated migration equations. The stronger effect in the local quit equations is due in part to the larger mean value for local quits[14] which is then applied to the logit coefficients to estimate marginal effects (see

[14]The mean value for *NMQUIT* is at least twice as large as the mean value for *MIGQT* in all three samples.

TABLE 4—DETERMINANTS OF 1966–71 MIGRATION: *NLS* OLDER MEN[a] MAXIMUM LIKELIHOOD LOGIT ESTIMATES

	Uncond	*Migqt*	*Miglay*	*Migtr*	*Nmquit*	*Nmlay*
EDUC	0072	0038	−0010	.0052	.0029	−.0064
	(3.85)	(3.24)	(−1.02)	(4.58)	(1.15)	(−2.38)
REM	0017	.0010	.0014	−.0007	.0005	.00007
	(1.33)	(1 29)	(2 21)	(−.91)	(.32)	(.04)
WAGE	−.0032	−0114	.0015	−.0003	−0229	.0030
	(−1.01)	(−3 40)	(1 49)	(−.19)	(−3.70)	(.73)
MAR	−.0131	−.0074	−.0056	.0198	−.0326	−.0034
	(−.64)	(−.62)	(−.60)	(.95)	(−1.09)	(−.10)
SCHL	−.0213	−.0090	−.0110	−.0013	.0092	−.0265
	(−1.71)	(−1.09)	(−1.63)	(−.19)	(.56)	(−1.52)
WLFP	−.0192	−.0065	−.0118	−.0017	.0385	−.0086
	(−1.34)	(−.75)	(−1.35)	(−.17)	(1.89)	(−.43)
WW	.0012	.0029	.0001	−.0019	−.0006	.0026
	(.38)	(1.89)	(04)	(−.53)	(−.10)	(.51)
HLTH	−.0110	.0013	−.0083	.0016	−.0239	.0401
	(−.79)	(.15)	(−1.25)	(.17)	(−1.31)	(1.81)
UNEMP	.0083	−.0045	.0128	−.0141	.0934	.1454
	(.40)	(−.32)	(1.61)	(−.70)	(4.44)	(6.84)
χ^2	23.51	25.26	16.58	29.52	50.51	58.68
N	1790	1790	1790	1790	1790	1790

[a]Asymptotic *t*-ratios are given in parentheses. Variables are defined in Table 2.

fn. 13). It can therefore be concluded that the relationship between wages and migration is strongly dependent on the fact that job separations accompany migration; the only case in which a move is seen to be negatively related to the wage (i.e., quitting and migrating) is found to be due entirely to the negative effect of the wage on the job separation itself.

B. *Education*

Education has a positive and significant effect in all samples on the unconditional probability of migrating. This result is consistent with the findings of other studies surveyed in Greenwood (1975) and has been explained as being due to the more educated individual's ability to adapt to new locations and his greater efficiency in searching for jobs in other locations. The empirical results in this paper show that although education is not positive and significant in all of the joint-probability migration equations, it does have an effect on migration that is independent of its effect on the probability of separating. This can be seen by comparing the coefficients in the local and nonlocal separation equations; in all cases *EDUC* is more positive in the case of a nonlocal separation. Unlike the wage, education does have an independent effect on the decision to migrate.

C. *Family Variables*

The costs of migration that are usually associated with marital status can be measured by information on the wife's labor force participation and the ages of the children. The effect of the wife's labor force status is measured by a dummy variable indicating her participation in the labor force (*WLFP*) and a continuous variable measuring her wage or annual earnings (*WW*, *WINC*).[15] For men in their 30's, 40's, or 50's (Tables 3 and 4) we find that wife's labor force participation has a negative but insignificant effect on the unconditional probability of migration.[16] This occurs for two reasons. First, wife's participation has no effect in these samples on the probability of being transferred. Second, although wife's participation does inhibit migration in the case of job quitters, this effect can not be directly observed in the joint-probability (*MIGQT*) equations. The reason is that wife's participation has a positive and significant effect on the probability of quitting locally. Therefore, to measure the negative effect of *WLFP* on the probability of migrating, one should compare the coefficients in the *MIGQT* and *NMQUIT* equations. As can be seen from Tables 3 and 4, *WLFP* is less positive in the *MIGQT* equations, reflecting the inhibiting effect of this variable on quitting into another labor market.[17] Similarly, the negative effect of school-age children (*SCHL*) on the migration decision of job quitters in Table 4 is shown by comparing the effects of this variable on *MIGQT* and *NMQUIT*. This analysis therefore shows the importance of decomposing the unconditional probability of migration in order to correctly estimate the extent to which a working wife and the presence of school-age children inhibit a job quitter from changing locations.

D. *Job Tenure and Length of Residence*

Previous research on migration has found that one of the most important determinants of the decision to migrate is the length of

[15] In the case of the Coleman-Rossi sample, many men whose wives worked failed to report their wages or earnings. Since there were so many missing values for this variable, it was deleted from the regressions for this sample.

[16] In the case of men who are in their 20's (Table 2), the effect of wife's labor force participation on the unconditional probability of migration depends on the amount of the wife's earnings. If the wife's annual earnings are below $5,800 her participation does not inhibit migration, at earnings levels above $5,800 her participation has a negative effect which eventually becomes significant. (This is calculated by realizing that *WINC* is actually an interaction term between the dummy variable *WLFP* and the wife's earnings if she works.) Since these women are in their childbearing years and are likely to participate intermittently in the labor force, their current participation is not an inhibiting factor in migration unless their earnings represent a substantial contribution to family income.

[17] This result also holds for the young men in Table 2.

residence in the current location.[18] Individuals who have lived in the current location a long time may be less likely to move because they have built up a stock of capital that is specific to this location; that is, over time, strong community ties will have been developed thereby raising the costs of migration. One must also recognize, however, that the negative effect of length of residence may be due to the relationship between residence and job tenure. To the extent that the individual has not changed jobs during his stay in this location, job tenure and length of residence will be strongly correlated. Moreover, there is substantial evidence that job tenure reduces the probability of a job separation because of the positive correlation between tenure and job-specific training.[19] Since the relationship between migration and job separation has already been documented, it is quite possible that the observed negative effect of length of residence on migration may be due to the negative effect of tenure on job separations. We would like to be able to identify whether there is an independent effect of residence on migration.

Fortunately, since the data sets provide information on both length of residence and job tenure, the separate effects of length of residence and job tenure on the decision to migrate can be identified. This is accomplished by defining a variable *RTEN* which equals the difference between length of residence and job tenure and including this variable as well as *JOB* (length of job tenure) in the regressions. Then *RTEN* captures the effect of a year of residence net of job tenure, that is, the "pure" residence effect, while the coefficient on *JOB* is the sum of the pure residence effect and the pure job effect, if it exists. If *RTEN* has a significant effect on the decision to migrate and the coefficient on *JOB* exceeds (in absolute value) the coefficient on *RTEN*, then it can be concluded that the inhibiting effect of residence observed in other studies is due to the acquisition of *both* location-specific capital and job-specific capital.

Table 5 contains the estimated coefficients on *JOB* and *RTEN* for each of the unconditional and joint-probability equations. The results show that *RTEN* has a significant negative effect in the equations referring to the young cohorts (*NLS* Young Men and Coleman-Rossi); thus for these samples length of residence has an inhibiting effect on migration which is net of the relationship between residence and tenure. For the older men, however, the correlation between residence and job tenure is very high; *RTEN* only ranges from zero to nine years. For these men, *RTEN* is insignificant indicating that the negative relationship between residence and migration is due solely to the negative effect of job tenure on separations. It is important to note, however, that for all three samples,

[18] For example, see Kaluzny and Polachek and Horvath. Recall that this observed effect may be due in part to serial correlation in the dependent variable.

[19] See the papers by Borjas and myself and by Jovanovic and Mincer.

TABLE 5—MAXIMUM LIKELIHOOD LOGIT COEFFICIENTS AND ASYMPTOTIC t-RATIOS ON *RTEN* AND *JOB* FOR ALL SAMPLES[a]

Dependent Variable	*RTEN*	*JOB*
NLS Young Men		
Unconditional	– 0088	–.0239
	(–8 16)	(·4 31)
Migrate and quit	–.0044	–.0244
	(–5.47)	(–5.01)
Migrate and be laid off	– 0018	–.0154
	(–3.81)	(–3.92)
Transfer	–.0027	.0029
	(–4.05)	(1.00)
Coleman-Rossi		
Unconditional	·.0240	–.0204
	(–5.76)	(–3.75)
Migrate and quit	–.0086	–.0100
	(–2.94)	(–2.42)
Migrate and be laid off	–.0022	– 0054
	(–1.43)	(–1.79)
Transfer	–.0156	–.0072
	(–2.21)	(–2.21)
NLS Mature Men		
Unconditional	–.0016	–.0042
	(–.80)	(–5.86)
Migrate and quit	–.0006	–.0042
	(–.45)	(–5.03)
Migrate and be laid off	–.0002	–.0013
	(–.15)	(–3.09)
Transfer	–.0011	–.0002
	(–.85)	(–.73)

[a] t-ratios shown in parentheses.

TABLE 6—COEFFICIENTS ON MIGRATION DUMMY VARIABLES FROM WAGE GROWTH REGRESSIONS[a]

	NLS Young Men 1971–73			Coleman-Rossi 1964–69			*NLS* Mature Men 1966–71		
GEOG	.3621 (2.32)			–.9982 (–.03)			–.0095 (–.06)		
MIGQT		.4820 (2.27)	.5341 (2.44)		–39.76 (–1.16)	–53.14 (–1.50)		–.0849 (–.34)	–.1502 (–.59)
MIGLAY		–.3147 (–.82)	–.2572 (–.67)		–129.80 (–1.97)	–143.10 (–2.16)		–.1628 (–.57)	–.2233 (–.77)
MIGTR		.4338 (1.89)	.4792 (2.05)		119.17 (2.73)	106.92 (2.42)		.1683 (.69)	.1327 (.54)
NMQUIT			.1379 (.85)			–36.14 (–1.80)			–.2838 (–2.17)
NMLAY			.1514 (.63)			–4.39 (–.41)			–.1059 (–.83)

[a]Definitions of variables are *GEOG* equals one if individual migrated during the period; *MIGQT* equals one if individual quit and migrated; *MIGLAY* equals one if individual was laid off and migrated; *MIGTR* equals one if individual migrated but did not change employers; *NMQUIT* equals one if individual quit but did not migrate and *NMLAY* equals one if individual was laid off but did not migrate.

when a job separation accompanies a geographic move, job tenure *itself* reduces the probability of migration, that is, the coefficient on *JOB* exceeds that on *RTEN*, because of the relationship between tenure and job separations. In the case of a transfer, however, the pure job effect (coefficient on *JOB* minus coefficient on *RTEN*) is actually positive; employers appear to be more likely to "promote" those individuals who have shown a commitment to the firm. This analysis therefore shows that the negative effect of residence on migration observed in other studies is misleading in two respects. First, when a job separation accompanies a geographic move, part of the inhibiting effect of residence is due to the negative effect of tenure on separations. Second, when the geographic move is an intrafirm transfer, the effect of residence may be nonnegative if tenure is not held constant since job tenure increases the probability of a transfer.

IV. Wage Gains from Migration

Previous work on migration has not conclusively established that migrants have larger wage gains than individuals of similar characteristics who do not migrate.[20] Since this article has already shown that it is important to distinguish among types of moves in examining the determinants of migration, the distinction may also help in obtaining a more accurate measure of the return to migration. Table 6 contains coefficients on dummy variables measuring migrant status from regressions on absolute wage growth for each of the three samples.[21] The migrant status dummy variables are defined in the footnote to the table. While a vector of standardizing variables was included in the wage growth regressions, these coefficients are not reported here.[22]

[20]Greenwood (1975) shows that while many studies have found a positive return to migration, others have been unable to support this conclusion. He argues that the return to migration can be correctly calculated only if the migrant population is disaggregated as finely as possible; in other words, the return to migration differs appreciably across groups. This paper suggests that job mobility may be an important characteristic by which migrants should be stratified.

[21]By analyzing the effects of migration on wage growth rather than wage levels, we avoid the possibility for simultaneity bias in the wage equation. If there are certain unobserved personal characteristics which affect both an individual's wage and his decision to migrate, a wage level equation will be biased. A wage growth equation nets out these unobserved individual differences which affect an individual's earnings throughout the life cycle.

[22]The vector includes education, years of experience, marital status, wife's labor force status and income, presence of school children, job tenure, length of residence, and unemployment experience.

The results in Table 6 show that if no distinction is made among types of moves, a positive and significant effect of migration on wage growth is observed only for the *NLS* Young Men. Distinguishing among moves related to quits, moves related to layoffs, and transfers provides a more revealing picture of the returns to migration. For men in their 20's and 30's (*NLS* Young Men and Coleman-Rossi), transfers have a positive and significant effect on wage growth. In other words, young men who are transferred by their employers achieve wage gains that are substantially larger than the gains of men with similar characteristics who do not migrate. It therefore appears that in this age group a transfer acts as a promotion within the firm. While men in their 50's who are transferred do not receive wage gains that are significantly larger than that of the nonmigrants, it is important to note that this type of move results in the largest wage gain (the coefficient on *MIGTR* is larger than those on *MIGQT* and *MIGLAY*).

Of all the coefficients on the separation-related moves, only one is significant: the *NLS* Young Men who quit and migrate achieve significantly larger wage increases than nonmigrants. Does this imply that for the two older cohorts a geographic move that accompanies a job separation does not pay? The answer to this question depends on with whom the migrant is being compared. For example, in all three samples, individuals who quit and migrate do better than individuals who are laid off and migrate (compare *MIGQT* and *MIGLAY*).[23,24] Further, in the *NLS* Mature Men sample, individuals who quit and migrate achieve larger wage gains than individuals who quit but do not migrate (compare *MIGQT* and *NMQUIT*). In general, however, one can conclude that of the three types of moves, transfers result in the largest payoffs.

[23]Note that for layoffs this does not hold; individuals who are laid off do better if they do not migrate.

[24]The reader may be puzzled as to why a significant positive return to local quitting is not observed. The paper by Borjas and myself shows that for the *NLS* Mature Men, a substantial proportion of the quits in this age group either result in increased job satisfaction but not increased money wages or are caused by exogenous factors such as health or family problems. For the young men, only those individuals who said they quit because they found a better job had significantly larger wage gains in the 1971–73 time period. Individuals who quit because of personal problems or because of dissatisfaction with their current jobs did not have larger wage gains than stayers.

V. Summary and Conclusions

This article has analyzed the determinants and consequences of migration at different stages in the life cycle. The theme of the article has been that migration is closely related to job mobility (in fact, between one-third and one-half of all moves are *caused* by the decision to *change* jobs) and that when the decision to migrate or the returns to migration are explored, one must take account of this relationship. Several findings support this argument:

1) Economic theory predicts that, *ceteris paribus*, the wage should have a negative effect on the decision to migrate. This article shows that the wage has a significant negative effect only in the case of the joint probability of migrating and quitting. Moreover, this negative coefficient is entirely due to the negative effect of the wage on the job separation itself.

2) The true inhibiting effect of a working wife on a man's decision to migrate is shown to be correctly estimated only when the unconditional probability of migrating is decomposed; this occurs because of conflicting effects of this variable on the transfer decision, the decision to change jobs in the local market, and the decision to quit and migrate. Similar problems exist for measuring the effect of the presence of school-age children.

3) Previous research on migration has found that one of the most important determinants of the decision to migrate is the length of residence in the current location. This article shows that since residence and job tenure are positively correlated, the effect of residence on migration is at least partially due to the relationship between job tenure and the decision to change jobs.

4) The wage gains from migration are

also seen to depend on the nature of the move and the age of the migrant. Of the three types of moves, transfers in general lead to the largest wage gains; this effect is significant, however, only for the two younger cohorts. A quit-related move is also found to lead to larger payoffs than a layoff-related move for all three samples.

In conclusion, the empirical findings presented in this article support the initial argument that one must take account of job mobility in studying the determinants and consequences of the decision to migrate. The results indicate that there is an important link between the decision to migrate and the probability of a job separation; an analysis of migration that ignores this link may fail to understand the role played by many socioeconomic variables in the migration process.

REFERENCES

A. Bartel and G. Borjas, "Middle-Age Job Mobility: Its Determinants and Consequences," in Seymour Wolfbein, ed., *Men in the Pre-Retirement Years*, Philadelphia 1977.

Z. D. Blum, N. L. Karweit and A. B. Sorensen, "A Method for the Collection and Analysis of Retrospective Life Histories," rept. no. 48, Johns Hopkins Univ. Center Stud. Soc. Organ. Sch. 1969.

S. Bowles, "Migration as Investment: Empirical Tests of the Human Investment Approach to Geographical Mobility," *Rev. Econ. Statist.*, Nov. 1970, *52*, 356–62.

J. DaVanzo, "Why Families Move: A Model of the Geographic Mobility of Married Couples," R-1972, Rand Corp., Sept. 1976.

M. Greenwood, "An Analysis of the Determinants of Geographic Labor Mobility in the United States," *Rev. Econ. Statist.*, May 1969, *51*, 189–94.

———, "Research on Internal Migration in the United States: A Survey," *J. Econ. Lit*, June 1975, *13*, 397–433.

R. Hall, "Why is the Unemployment Rate So High at Full Employment?," *Brookings Papers*, Washington 1970, *3*, 369–410.

B. Jovanovic and J. Mincer, "Labor Mobility and Wages," paper presented at Nat. Bur. Econ. Res. Conference on Low Income Labor Markets, Chicago, June 1978.

R. Kaluzny, "Determinants of Household Migration: A Comparative Study by Race and Poverty Level," *Rev. Econ. Statist.*, Aug. 1975, *57*, 269–74.

John B. Lansing and Eva Mueller, *The Geographic Mobility of Labor*, Ann Arbor 1967.

L. H. Long, "The Influence of Number and Ages of Children on Residential Mobility," *Demography*, Aug. 1972, *9*, 371–82.

Ira S. Lowry, *Migration and Metropolitan Growth*, San Francisco 1966.

J. Mincer, "Family Migration Decisions," *J Polit. Econ.*, Oct. 1978, *86*, 749–73.

D. Parsons, "Specific Human Capital: An Application to Quit Rates and Layoff Rates," *J. Polit. Econ.*, Nov./Dec. 1972, *80*, 1120–43.

S. Polachek and F. Horvath, "A Life Cycle Approach to Migration: Analysis of the Perspicacious Peregrinator," in Ronald Ehrenberg, ed., *Research in Labor Economics*, Vol. 1, Greenwich 1977.

A. Schwartz, "Interpreting the Effect of Distance on Migration," *J. Polit. Econ.*, Sept./Oct. 1973, *81*, 1153–69.

L. A. Sjaastad, "The Costs and Returns of Human Migration," *J. Polit. Econ.*, Oct. 1962, *70*, 80–93.

U.S. Department of Labor, *The Pre-Retirement Years*, Vols. 1–4, Manpower Res. Monog. No. 15, Washington 1970–74.

———, *Career Thresholds*, Vols. 1–6, Manpower Res. Monog. No. 16, Washington 1970–77.

Job Opportunities, the Offered Wage, and the Labor Supply of Married Women

By MASAO NAKAMURA, ALICE NAKAMURA, AND DALLAS CULLEN*

The major difference between segmented labor market and human capital theories about the labor force behavior of women lies in the attention paid to micro vs. market-wide or macro variables. In a study such as James Heckman's (1976) which includes no market variables, the demand for the labor of married women in any given education-experience (and hence offered wage) class is implicitly assumed to be infinitely elastic. Thus the observed differences in the labor force behavior of individual women are attributed entirely to differences in supply characteristics such as education and child status. On the other hand in segmented labor force analyses, such as Barbara Bergmann's and Irma Adelman's study, the macro phenomenon of occupational segregation by sex is seen as the major factor affecting the participation, wage rates, and hours of work of women. These two types of studies lead to different explanations of why the labor force participation of women has increased in recent years. Different sets of government policies aimed at improving the labor force situation of women are also implied.

In this paper we present a model of the labor force behavior of married women in which both individual and family decision making, and macro labor market conditions are found to play important roles. An unemployment variable and an index summarizing the ratio of expected available local job slots for women to the potential female labor force population are incorporated into a marginal utility analysis of the labor force behavior of married women in Canada. The inclusion of the local opportunity for jobs variable is supported by detailed evidence on the labor force segregation of women in Canada. Consistent estimation results are presented for eleven age groups in a probit analysis of whether or not a married woman works, and for eight age groups in equations estimating the offered wage rates and annual hours of work of married women who do work.

One unexpected finding is that working wives in Canada tend to work fewer hours per year when paid more per hour. This is contrary to the findings of other researchers for the United States, and has important policy implications. Although it is possible that our results differ from those of other researchers solely because we have analyzed data for another country, we argue in Section V of this paper that the difference in results is more likely due to differences in the form in which the labor supply function for wives is estimated and the choice of the variables which are used to control for child status. Our resulting uncompensated wage elasticities of hours of work are shown to be very similar to those reported by other researchers for men.

The data base used in this study is the Family File of the first Public Use Sample to be made available from a Canadian census. Combined grouped R^2s are presented showing the extent to which our equations explain the observed macro variations in the labor force behavior of married women classified by various characteristics. Finally we use our estimated model to see what changes we would expect in the labor force behavior of a hypothetical 41-year-old wife living in a small city in New Brunswick given a variety of changes

*Faculty of Business Adminstration and Commerce, University of Alberta For further computational results and theoretical arguments supporting various statements in this paper, see our book. The work for this paper was supported in part by the Statistics Canada-SSRCC Programme of 1971 Census Analytical Studies, and by the Faculties of Graduate Studies and Research and of Business Administration and Commerce of the University of Alberta. The empirical results in this paper are primarily based on Public Use Sample Data derived from the 1971 Canadian Census of Population supplied by Statistics Canada. The responsibility for the use and interpretation of these data is entirely ours. We would like to thank T. Daniel, K. Gupta, anonymous referees, and the managing editor for their helpful comments, and James Heckman for making available to us some of his work which had not yet been published.

in her situation brought on by herself, her family, the government, or by particular types of economic development in New Brunswick. Substantial wage gains, leading to higher expected annual income levels with fewer hours of work, are found to result from more education, from legislation or changes in social attitudes leading to less occupational segregation by sex, or from selective economic growth within the present structure of labor market segregation. Higher expected income levels with more hours of work are found to result from reduced childbearing, or from reductions in the costs of child care.

I. The Model

Looking at the 1971 sample data for Canadian families we find that despite substantial differences in the mean hourly wage of the husband and other characteristics of families with and without working wives, the differences in the mean annual numbers of hours worked by the husbands in these two groups of families are negligible. In attempting to describe the short-run (annual) labor force behavior of the wife, therefore, it may be useful to treat the husband's hours of work (and hence his earned income) as given. We will view the household as maximizing a twice-differentiable quasi-concave conditional utility function $U(x,l;Z^*)$ subject to the income and time constraints,

$$px = A + I + wh \tag{1}$$

$$T = l + h \tag{2}$$

and $0 \leq h \leq T$, where x is a Hicksian composite good representing the consumption of all goods other than leisure, l represents the nonmarket time (hours of leisure) of the wife, h $(= T - l)$ represents the market time (hours of work) of the wife at offered (market) wage w, T is the total time available, p is the price of the Hicksian composite good, A is asset income, and I is the annual income of the husband. Denote Z^* to be a vector of constraints arising from previous choices, such as the number of children and education. Assuming that h is strictly less than T, the Lagrangean for this problem, for any given vector Z^*, is

$$V = U(x, T - h; Z^*) + \gamma h + \lambda(A + I + wh - px)$$

and the Kuhn-Tucker conditions are (1), (2), $\gamma \geq 0$,

$$U_x - \lambda p = 0 \tag{3}$$

$$-U_l + \gamma + \lambda w = 0 \tag{4}$$

$$\gamma h = 0 \tag{5}$$

where $U_x = \partial U/\partial x$ and $U_l = \partial U/\partial l$. Then rearranging (1) and (3) we get

$$x = (A + I + wh)/p \tag{6}$$

$$\lambda = U_x/p = [\partial U(x, T - h; Z^*)/\partial x]/p \tag{7}$$

From these last two equations it is seen that, given any offered wage w, in equilibrium λ is a function of $h, p, A+I, wh$, and Z^*. From (4) we also have that in equilibrium $\lambda = (U_l - \gamma)/w$. Thus $w = (U_l/\lambda) - (\gamma/\lambda) = w^* - (\gamma/\lambda)$ where the shadow price of a wife's time (asking wage),

$$w^* = U_l/\lambda \tag{8}$$

depends on h, p, $A + I$, wh, and Z^* when $h > 0$ and[1] on p, $A + I$, and Z^* when $h = 0$ Since $\lambda > 0$ by (7), and since (5) implies $\gamma = 0$ if $h > 0$ and $\gamma \geq 0$ if $h = 0$, we have $w = w^*$ if $h > 0$ and $w \leq w^*$ if $h = 0$. One crucial aspect of this solution is that the wife's asking wage w^* depends on her income wh, which depends on her offered wage w. This is because the mechanism by which leisure is traded for the increased consumption of other goods is through the relaxation of the household budget constraint.

If we take the *log* of both sides of (8) and linearize it around Z_i^*, A_i, I_i, $\ln w_i$, and h_i, for the ith married woman we get

$$\ln w_i^* = \begin{cases} \beta_0 + Z_i^*\beta_1 + \beta_2 A_i + \beta_3 I_i + \beta_4 \ln w_i + \beta_5 h_i + \mu_i^* & \text{if } h_i > 0 \\ \beta_0 + Z_i^*\beta_1 + \beta_2 A_i + \beta_3 I_i + \mu_i^* & \text{if } h_i = 0 \end{cases} \tag{9}$$

where β_2 and β_3 are expected to be equal. The

[1] U_l is taken to be the left derivative of U with respect to l at $l = T$ (or at $h = 0$) since U_l is not defined for $l > T$ See also Heckman 1974, Appendix 1.

variable p does not appear in (9) because it is assumed to be the same for all households. Although the wife's offered wage w is also a price variable, it cannot be ignored in this way since it differs systematically from one wife to another. We will assume that variations in the wife's offered wage w are explained by

$$(10) \qquad ln\ w_i = \alpha_0 + Z_i\alpha_1 + E_i\alpha_2 + \mu_i$$

where Z and E are vectors of personal and regional economic variables, respectively.

It is not possible in our model to uniquely determine the coefficients of the shadow price equation (9). However, we can estimate the following expression for the wife's equilibrium number of hours of work:

$$(11) \quad h_i = \frac{1}{\beta_5}[(1-\beta_4)ln\ w_i - \beta_0 - Z_i^*\beta_1 - \beta_2 A_i - \beta_3 I_i - \mu_i^*] \qquad \text{at } h_i > 0$$

We will briefly discuss our choice of the variables defining the vectors Z^*, Z, and E.

II. Factors Affecting a Married Woman's Asking Wage (Z^*)

The greater the costs of working, the higher a wife's asking wage should be. As proxies for these costs we have included the number of children younger than 6, the number of children 6–14 years, and the product of these two variables. The interaction term is included to account for nonlinearities in the amount of time spent per additional child as both the number of children and the number of older children increase. In addition it is expected that the costs of a wife working will be higher the more hours she works.

The number of children ever born has been included as a proxy for basic feelings of couples about the advantages of family vs. market-oriented activities. We have also included a variable for whether or not the wife's religion is Roman Catholic and a variable for whether or not the wife lives in a French-speaking household.

A family's need for additional income should increase as income from other sources decreases and as financial obligations increase; and the wife's asking wage is expected to decrease as her family's need for additional income increases. We have included separate terms for the yearly employment income of the husband, the asset income of the family, and an interaction term constructed by adding these two income variables and dividing by the family size. The only variable acting solely as a proxy for financial obligations is the number of children 19–24 years attending school.

III. Factors Affecting a Married Woman's Offered Wage

The factors believed to affect the offered wage can be divided into two broad groups. The first consists of personal attributes of individual women. The second consists of characteristics of the labor market which affect the general levels of demand for workers of different types. Studies which incorporate variables from only the first of these two groups provide insight into the question of which women will work given the labor market conditions, but may be unable to explain variations in the overall level of the employment of women over time or from one locality to another.

A. *Personal Characteristics (Z)*

The personal characteristics considered in this study are similar to those included in previous studies, and are influenced by the availability of data. A wife's offered wage is expected to be higher the more education she has and the longer she waited to get married. The number of children younger than 6 is included as a proxy for the recentness of woman's potential job experience. Finally, rather than guessing at how age affects offered wages, we have carried out our analysis for several different age groups.

B. *Regional Economic Characteristics (E)*

When the unemployment rate is high more workers compete for each available job. Thus the offered wage rate should be driven downward. Similarly the offered wage rate should be lower in areas where there is a structural scarcity of jobs believed suitable for women.

TABLE 1—OCCUPATIONAL CHARACTERISTICS OF THE CANADIAN LABOR FORCE

Occupation	Women as Percent of Total Workers			Provincial Low and High Percentages of Female Workers	Percentage Growth of Total Workers	1971 Distributions of Total Workers by Residence		1971 Distributions of Workers		
	1951	1961	1971	1971	1951-1971	Urban	Rural	Total	Female	Married Female
Managerial	8.7	10.4	15.7	14.5-20.0	-17.1	4.9	2.0	4.3	2.0	1.9
Natural Sciences	6.9	4.8	7.3	2.1-9.4	250.8	3.1	1.2	2.7	0.6	0.5
Social Sciences	27.8	29.4	37.4	34.4-45.6	272.9	1.1	0.4	0.9	1.0	0.8
Religion	39.7	28.9	15.7	8.3-31.1	-23.0	0.3	0.3	0.3	0.1	0.0
Teaching	67.2[a]	64.4[a]	60.4[a]	57.8-71.5	200.4	4.2	3.5	4.1	7.1	7.4
Medicine	68.5[a]	72.1[a]	74.3[a]	71.1-78.2	193.9	4.2	2.1	3.8	8.2	7.9
Artistic	30.7	31.2	27.2	18.5-31.6	124.1	1.1	0.4	0.9	0.7	0.6
Clerical	56.1[a]	61.0[a]	68.4[a]	58.9-74.1	119.6	18.3	7.4	15.9	31.8	31.1
Sales	33.3[a]	32.0[a]	30.4[a]	25.2-38.5	165.2	10.4	6.0	9.5	8.4	9.0
Service	45.1[a]	46.7[a]	46.2[a]	37.2-54.9	91.8	11.9	8.8	11.2	15.1	14.2
Farming	3.9	11.7	20.9	12.5-24.2	-38.3	1.1	24.0	6.0	3.6	4.8
Other Primary	0.1	0.3	1.3	0.4-2.6	-28.8	1.1	4.5	1.8	0.1	0.0
Processing	14.8	13.7	17.8	11.3-43.3	-13.2	3.8	4.4	3.9	2.0	2.2
Machining and Fabricating	18.0[a]	17.9[a]	18.7[a]	3.0-25.2	32.1	10.8	8.0	10.2	5.5	6.3
Transport	0.5	0.6	2.4	0.8-3.5	22.2	3.7	4.7	3.9	0.3	0.3
Other	16.3	13.6	15.7	4.4-19.2	153.3	5.8	4.8	5.6	2.6	2.7
Unknown	20.6	26.0	43.4[a]	37.7-45.2	1044.5	8.4	9.4	8.5	10.8	10.0
All Occupations	22.0	27.3	34.3	27.6-35.8	62.8					

Source: Calculated from the *1951 Census of Canada*, Vol IV, Table 4, *1961 Census of Canada*, Vol. III, Part I, Table 6, *1971 Census of Canada*, Vol. III, Part 2, Tables 2 and 8

[a]More than 5 percent of the female labor force was in this occupation in the given year.

For instance, it has been argued that rural areas, and particularly rural nonfarm areas, have few jobs for women. In such areas the offered wage rate for women may normally be so low that many decide not to even look for a job.

Looking at females as a percent of total workers in each of the major occupational classifications, we find a great deal of consistency over time for Canada, and from one province to another in the same time period.[2] These percentages for 1951, 1961, and 1971 for Canada,[3] together with the provincial extremes for 1971, are shown in columns 1–4 in Table 1. For instance, we see that women generally have made up 40 percent or more of the total work force in the Teaching, Medi-

[2]The finding that women are occupationally segregated is not new. See Edward Gross, Valerie Oppenheimer, Abbott Ferris, and Bergmann and Adelman for *U.S.* studies, and Sylvia Ostry and Morley Gunderson for Canadian studies. Moreover Oppenheimer presents detailed evidence showing that the personal and family characteristics used as explanatory variables in most studies of the labor force behavior of women simply cannot account for the observed rise since World War II in their labor force participation.

[3]Occupational comparisons between 1951, 1961, and the 1971 Canadian censuses are difficult to make because of changes in the occupational classification scheme. We made these comparisons by going back to the dictionary definitions of each of the minor occupational codes for 1951 and 1961, and reordering these into the 1971 major occupational groupings. A complete description of how this reclassification was carried out is available on request from Cullen.

TABLE 2—RATIOS OF EXPECTED JOBS FOR WOMEN TO NUMBER OF WOMEN 15 YEARS OF AGE AND OLDER, BY PROVINCE AND PLACE OF RESIDENCE FOR 1971

	By Occupation			
	Urban		Rural	
Province	30,000 and Over	Under 30,000	Nonfarm	Farm
Newfoundland	.45	.34	.20	.19
Nova Scotia	.50	.38	.29	.30
New Brunswick	.48	.39	.28	.29
Quebec	.42	.35	.26	.17
Ontario	.49	.41	.35	.26
Manitoba	.50	.41	.32	.18
Saskatchewan	.48	39	.27	.15
Alberta	51	43	.32	.17
British Columbia	.45	40	.36	.33

Source Calculated from *1971 Census of Canada*, Public Use Sample Tape–Individual File

cine, Clerical, and Service occupations, and 10 percent or less of the total work force in the Natural Sciences, Other Primary, Construction, and Transport occupations. Moreover, from column 5 we find that in general women are represented most heavily in the occupations which have grown more rapidly than the mean rate for all occupations.[4]

In 1951, 61.6 percent of all people and 64.8 percent of all married women in Canada lived in urban places of residence. The comparable figures for 1971 are 76.1 and 77.9 percent, respectively. Columns 6 and 7 of Table 1 show that, compared with rural areas, urban areas have higher percentage concentrations of workers in occupations which employ larger percentages of women. Thus, along with the growth over time in job opportunities for women, there has also been a shift in the percent of the population living in places offering relatively more job opportunities for women. Table 2 shows the expected numbers of jobs per woman, by province and place of residence. The expected numbers of jobs for women in each occupation, in each province and place of residence, are calculated by multiplying the 1971 Canada-wide percentages of women in each occupation by the actual number of workers in each occupation, in each province and place of residence. These expected numbers are summed over all occupations in each province and place of residence, and the resulting totals are then divided by the total number of women 15 years and older in each province and place of residence.

In our study each married woman is assigned the average 1970 unemployment rate for her province, and an appropriate value from Table 2 of the local opportunity for jobs index. A rural area dummy variable has also been included.[5]

IV. Estimation of the Model

The equations of interest in our study are

$$(10) \quad ln\ w_i = \alpha_0 + Z_i\alpha_1 + E_i\alpha_2 + \mu_i$$

$$(11') \quad h_i = \frac{1}{\beta_5}\,[(1-\beta_4)ln\ w_i - \beta_0 - Z_i^*\beta_1 - \beta_2 A_i - \beta_3 I_i] + v_i^*$$

[4]This finding is not new either. See Victor Fuchs, p. 237

[5]The above discussion concerns the occupational segregation of all women, not married women. This is because comparable figures for workers classified by marital status as well as sex are not available for 1951 and 1961. However, from the evidence presented for 1971 in columns 8-10 of Table 1, it would appear that our generalizations about the occupational segregation of all women are equally valid for married women considered by themselves. Subsequent to developing our opportunity for jobs index we found that William Bowen and T. Aldrich Finegan (pp. 772–76) calculated a similar index for the United States which differs from ours in that the denominator of their index for each geographical region is the total civilian employment, rather than the potential female labor force as in the case of our index For purposes of examining the wage rate and labor force behavior of married women we feel that our index is more appropriate, although use of either index probably would represent an improvement compared with the common practice in cross-sectional labor force studies of ignoring labor market conditions.

where $v_i^* = -(1/\beta_5)\mu_i^*$. We assume the covariance structure:

$$E(\mu_i) = E(v_i^*) = 0$$

$$E(\mu_i, \mu_j) = \begin{cases} \sigma_1^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$E(\mu_i, v_j^*) = \begin{cases} \sigma_{12} & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

$$E(v_i^*, v_j^*) = \begin{cases} \sigma_2^2 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$$

We observe values of w_i only for those women who worked for pay in 1970 (i.e., $h_i > 0$). Hence we need to include terms in (10) and (11′) to correct for selection bias, as suggested by Heckman (1976). (See also Reuben Gronau, 1973, 1974, and H. Gregg Lewis.) Assuming a joint normal distribution of μ_i and v_i^* the selection biases for (10) and (11′) are $E(\mu_i | h_i > 0) = (\sigma_{12}/\sigma_2)\lambda_i$ and $E(v_i^* | h_i > 0) = \sigma_2\lambda_i$, respectively, where $\lambda_i = f(\phi_i)/F(\phi_i)$ and $f(\phi)$ and $F(\phi)$ are the density and cumulative density functions of the standard normal distribution. The denominator of λ_i is the probability that observation i has data for the offered wage rate w_i. Moreover the lower the probability that an observation has data for w_i, the greater the value of λ for that observation. (See Heckman, 1976, p. 479.) The ϕ_i, and hence the λ_i, are derived by probit analysis as follows. From our maximization problem we have

$$P(D_i = 1) = P(h_i > 0) =$$
$$P(ln\ w_i - ln\ w_i^* |_{h_i = 0} > 0) =$$
$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\phi_i} e^{(-t^2/2)}\, dt$$

where D_i is defined to be one or zero depending on whether or not the ith married woman works. From equations (9) and (10) it can be shown that

$$(12)\quad \phi_i = \frac{1}{\sigma}[(\alpha_0 - \beta_0) + Z_i\alpha_1 - Z_i^*\beta_1 + E_i\alpha_2 - \beta_2 A_i - \beta_3 I_i]$$

where σ^2 is the variance of the random term $\mu_i - \mu_i^*$. Probit analysis gives estimates for the coefficients $(\alpha_0 - \beta_0)/\sigma$, α_1/σ, $-(\beta_1/\sigma)$, α_2/σ, $-(\beta_2/\sigma)$, and $-(\beta_3/\sigma)$, which in turn can be used to calculate ϕ_i and λ_i for the ith married woman.

After estimating the probit coefficients from the entire sample of married women, we used a two-stage generalized least squares (*GLS*) procedure[6] to estimate the following regression equations for $ln\ w_i$ and h_i using the subsample of married women who actually worked:

Offered wage equation:

$$(13)\quad ln\ w_i = \alpha_0 + Z_i\alpha_1 + E_i\alpha_2 + (\sigma_{12}/\sigma_2)\lambda_i + V_i$$

Hours equation:

$$(14)\quad h_i = \frac{1}{\beta_5}[(1 - \beta_4) ln\ w_i - \beta_0 - Z_i^*\beta_1 - \beta_2 A_i - \beta_3 I_i] + \sigma_2\lambda_i + V_i^*$$

where the covariance structure of V_i and V_i^* is shown to be $E(V_i^{*2}) = \sigma_2^2 M_i$, $E(V_i^2) = \sigma_1^2 (1 - \rho^2) + \sigma_1^2\rho^2 M_i$, $E(V_i^*, V_i) = \sigma_{12} M_i$, $M_i = 1 + \phi_i\lambda_i - \lambda_i^2$, and $\rho = \sigma_{12}/\sigma_1\sigma_2$ (see Heckman, 1976).

The basic data for this study consists of the

[6]Since the disturbance terms V_i and V_i^* have heteroskedastic variances over a cross section of married women and since $ln\ w_i$ appears in (14) we used the following combination of two-stage least squares and *GLS* to estimate (13) and (14): (i) Use *OLS* to estimate the coefficients of (13). Then the variance of V_i is estimated as $a + bM_i$ where a and b are the estimated intercept and slope when the squares of the *OLS* residuals from (13) are regressed on M_i which is calculated for each married woman who worked. The *GLS* estimates are then found by applying weighted *OLS* to (13) using the estimated variances of the disturbance term, and predicted values are calculated for $ln\ w_i$. (ii) The predicted values calculated for $ln\ w_i$ in (i) are substituted for the actual values of $ln\ w_i$ in (14) and *OLS* estimates are obtained for this relationship. Then the variance of V_i^* is estimated as $s_2^2 M_i$ where s_2 is the *OLS* estimate of the coefficient of λ_i in (14) obtained from the first iteration, and *GLS* estimates are found for (14) by using weighted *OLS* with these estimated variances. In deriving *GLS* estimates for (13) and (14) the interequation correlation was ignored. The standard errors reported in this paper might be downward biased (see Heckman, 1977), but our substantive results appear to hold unchanged. Although expressions for adjusted standard errors are presented for both *OLS* and *GLS* estimates in Heckman (1977), the *GLS* expressions appear to be computationally intractable. Heckman himself resorts to *OLS* estimates in this paper despite his previous arguments (1974, 1976) about the importance of heteroskedasticity in his model.

40,665 records for married couples living in Canada in households with no nonrelatives present which are contained in the 1 percent Family Public Use Sample from the *1971 Census of Canada*. These records were divided into eleven groups according to the age of the wife: 15–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, and 65+. Estimation results for the probit coefficients and for the offered wage and hours of work equations are presented in the following section.

V. Empirical Results

Since our local opportunity for jobs index variable has not been used in other studies, we first estimated the probit coefficients for our model with and without this variable for the age groups 15–19, 20–24, and 25–29. The only significant differences between these two sets of estimates are that when the jobs index variable is included the coefficients for the rural area dummy decrease in absolute magnitude and are never significant at a

TABLE 3—PROBIT ESTIMATES FOR FINAL MODEL[a]

Explanatory Variables	Age Groups										
	15-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54	55-59	60-64	65+
Constant	-6.750 (6.18)	-1.541 (4.75)	-1.075 (4.18)	-.469 (2.01)	-.541 (2 57)	-.627 (3 14)	-1.073 (5 50)	-1.507 (7.46)	-1.762 (7.64)	-1.809 (5.83)	-1.658 (4.65)
Education	.016 (.52)	.054 (6 05)	.088 (12.15)	.059 (8.47)	.062 (8.90)	.058 (8 29)	.065 (9 27)	.062 (7 88)	077 (9.15)	.064 (6.65)	.055 (5.34)
Number of children younger than 6	-.151 (.96)	-.474 (8 84)	-.381 (7 56)	- 456 (9 47)	- 420 (8.33)	- 406 (6.71)	- 574 (5.50)				
Number of children 6-14		-.142 (1 05)	- 130 (2 53)	-.126 (3.61)	-.096 (3 78)	-.157 (6 59)	-.151 (5.78)	- 094 (2.80)	-.109 (1.44)		
Product of number of children younger than 6 and number 6-14		.172 (1 73)	.119 (4.94)	.028 (1 61)	.046 (2.68)	.064 (2 89)	146 (3 74)				
Number of children 19-24 attending school				-.434 (88)	317 (2 64)	-.010 (.21)	.082 (2.07)	066 (1.41)	.253 (3.98)	.314 (2.49)	-.506 (.95)
Total of children ever born			- 102 (2.94)	.001 (.04)	.003 (.13)	- 008 (.48)	- 015 (1.09)	-.010 (.72)	-.001 (.10)	-.008 (.53)	-.031 (1.93)
Husband's income	- 00008 (85)	-.00016 (6 52)	- 00011 (8 32)	- 00008 (7.32)	-.00005 (6.26)	-.00003 (4.42)	- 00004 (6.19)	- 00003 (3 85)	-.00006 (4.56)	-.00001 (.54)	.00005 (1.60)
Asset income of family	- 00032 (1.74)	-.00011 (2 15)	-.00009 (4.01)	-.00007 (2.43)	- 00002 (91)	- 00001 (.48)	-.00004 (2.70)	-.00003 (2.11)	-.00007 (3.99)	-.00007 (2 89)	-.00004 (1.13)
Per person family income excluding wife's earnings)	00038 (1 78)	00037 (5.67)	00022 (5 74)	.00015 (3.71)	00007 (2 12)	- 00001 (.46)	00003 (1.38)	.00002 (87)	.00007 (2.33)	.00003 (.66)	-.00001 (.18)
Age at first marriage	.310 (5.32)	.057 (4.26)	.019 (2.15)	.006 (.78)	-.011 (1.72)	-.008 (1.61)	.002 (54)	.012 (2.79)	.005 (1.09)	.004 (.57)	-.012 (1.88)
Religion dummy	-.063 (.47)	034 (.63)	.133 (2.87)	.148 (3.04)	.091 (1.85)	.070 (1.41)	- 023 (.44)	-.112 (1 82)	.026 (.39)	.033 (.39)	.098 (.99)
Language dummy	.118 (.62)	-.190 (2.79)	-.282 (4.81)	-.026 (4.25)	-.380 (5.83)	-.288 (4.46)	-.280 (4.05)	-.188 (2.42)	-.348 (3.93)	-.383 (3.43)	-.102 (.77)
Provincial unemployment rate	-.044 (1.17)	-.024 (1.62)	-.007 (.54)	-.036 (2.70)	-.028 (2.04)	-.032 (2.38)	-.024 (1.77)	-.024 (1.65)	-.015 (.94)	.011 (.55)	-.025 (1.05)
Opportunity for jobs index	3.011 (4.42)	2.154 (7.59)	1.560 (6.23)	1.442 (5.72)	1.761 (7.21)	2.107 (8.70)	2.199 (9 00)	2.031 (7.73)	2.103 (7.06)	1.216 (3.20)	1.114 (2.50)
Combined Grouped R^2 = .9072[b]											
Pseudo R^2	.20	.26	.24	.15	.12	.11	.12	.09	.09	.06	.03
Maximum R^2 for model	.75	.73	.75	.74	.74	.74	.74	.74	.72	.66	.43
Pseudo R^2 for model (Pseudo R^2 divided by maximum R^2 for model)	.28	.36	.32	.21	16	.15	.17	.12	.13	.09	.08
Number in sample	607	4438	5541	4762	4613	4570	4476	3509	2941	2159	3049
Number who worked	328	2853	2651	1879	1755	1896	1900	1352	965	492	252
Proportion who worked	.54	.64	.48	.40	.38	.41	.42	.38	.33	.23	.08
Final value of log of likelihood function	-349	-2213	-3084	-2803	-2774	-2825	-2753	-2181	-1718	-1095	-816

Source: Calculations based on *1971 Census of Canada*, Public Use Sample Tape-Family File.

[a]Numbers in parenthese are (asymptotic) t-statistics. A coefficient is significant with at least a 95 percent level of confidence if its t-statistic is greater than or equal to 1.96.

[b]Explained in the text.

TABLE 4—GENERALIZED LEAST SQUARES ESTIMATES FOR *Log* OF OFFERED WAGE EQUATION[a]

Explanatory Variables	Age Groups							
	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59
Constant	-.957 (4.25)	.331 (1.78)	-.058 (.25)	-.692 (1.99)	1.310 (5.69)	-1.228 (4.99)	-1.280 (2.36)	-1.632 (2.95)
Education	.039 (7.65)	.070 (14.98)	.070 (11.53)	.066 (9.97)	.045 (8.39)	.067 (12.28)	.060 (7.81)	.077 (8.39)
Number of children younger than 6	.011 (.20)	-.002 (.07)	.066 (1.70)	-.080 (1.83)	.026 (.45)	-.257 (2.59)		
Age at first marriage	.027 (4.68)	.008 (1.72)	.003 (.70)	.009 (1.79)	-.003 (.91)	.004 (1.24)	.004 (1.18)	.011 (2.82)
Provincial unemployment rate	.021 (2.53)	.001 (.15)	.005 (.41)	.004 (.30)	.013 (1.06)	-.005 (.44)	-.004 (1.66)	.009 (.55)
Opportunity for jobs index	.386 (2.15)	-.283 (1.56)	.466 (2.00)	.604 (2.40)	.279 (1.45)	1.135 (5.30)	1.334 (4.19)	1.192 (3.53)
Selection bias	.136 (.83)	-.235 (2.38)	-.177 (1.13)	.263 (1.49)	-.894 (4.86)	.659 (4.32)	.665 (2.46)	.409 (1.63)
Combined grouped R^2 = .7307[a]								
R^2	.20	.34	.21	.11	.15	.22	.10	.22
Standard error of regression	1.01	1.00	1.01	1.00	1.03	1.01	1.00	1.00

Source. See Table 3.
[a]See Table 3

confidence level of even 80 percent, while the coefficients for the unemployment rate variable also decrease in both absolute size and significance. Thus we decided to carry out the rest of our study using a model in which the jobs index variable is included but the rural area dummy is not.

The probit coefficients for our final model are shown in Table 3 for all eleven age groups. Looking only at those coefficient estimates which are significant with a 95 percent level of confidence, we see that our hypotheses as summarized in Sections II and III of this paper are generally supported. Notice that the coefficients for the employment income of the husband and for the family asset income are negative, while the coefficients for the per person family income variable are generally positive. The implication is that the decline in the probability of a wife working as the income of the husband becomes larger, or the family asset income becomes larger, will be less steep the smaller the number of children. The only surprising finding in Table 3 is that Roman Catholic wives with childbearing patterns similar to other wives in their age group have higher than average probabilities of working.[7] Only wives living in French-speaking homes have below average probabilities of working, even after controlling for childbearing patterns.

Our estimation results for equations (13) and (14) are shown in Tables 4 and 5.[8] All coefficients significant at a 95 percent level have the expected signs, with the exception of the coefficient of the provincial unemployment rate in the offered wage equation. This may be due to the fact that wage rates for existing job slots are quite insensitive to unemployment rates. Also it may be that local

[7]This result is not believed to be due to multicollinearity with the dummy variable set equal to 1 if the language of the home is French. For the eleven age groups the correlation between these two variables are .51, .56, .57, .58, .59, .64, .65, .65, .67, and .70, respectively.

[8]The age groups 15–19, 60–64, and 65+ were dropped from this portion of the analysis due to the small numbers in these age groups of wives who actually worked, and other difficulties encountered in obtaining statistically meaningful results.

TABLE 5—TWO-STAGE GENERALIZED LEAST SQUARE ESTIMATES FOR ANNUAL HOURS OF WORK EQUATION[a]

Explanatory Variables	Age Groups							
	20–24	25–29	30–34	35–39	40–44	45–49	50–45	55–59
Constant	1899.474 (13.88)	2308.443 (11.99)	1829.135 (6.75)	2022.421 (8.21)	2313.827 (8.08)	1655.294 (6.07)	1410.875 (2.49)	1582.385 (2.68)
Log of wife's offered wage	-263.476 (2.13)	-390.432 (3.17)	-202.650 (1.30)	-232.312 (1.56)	-394.536 (2.38)	124.296 (.69)	393.912 (1.11)	154.818 (.59)
Number of children younger than 6	-202.806 (4.02)	-196.178 (3.90)	-182.250 (2.25)	-202.102 (3.18)	-236.485 (3.79)	-188.922 (1.82)		
Number of children 6–14	-74.928 (.88)	-28.094 (.70)	-15.128 (.42)	-32.803 (1.24)	-57.749 (2.18)	-67.800 (1.91)	-50.518 (1.27)	-43.236 (.50)
Product of number of children younger than 6 and number 6–14	106.312 (1.75)	-2.127 (.11)	-15.201 (.98)	29.129 (1.84)	81.746 (3.99)	10.106 (.25)		
Number of children 19–24 attending school			-567.839 (1.24)	57.836 (.53)	15.070 (.35)	-5.737 (.15)	23.066 (.48)	37.649 (.46)
Number of children ever born		-3.502 (.13)	15.839 (.60)	-16.246 (.95)	-14.643 (1.09)	-6.616 (.54)	-21.621 (1.53)	-11.760 (.81)
Husband's income	-.038 (2.38)	-.030 (2.49)	-.062 (4.84)	-.019 (2.17)	-.022 (3.01)	-.030 (3.72)	-.040 (3.46)	-.025 (1.31)
Asset income of family	-.044 (1.57)	-.040 (2.10)	-.042 (1.78)	-.009 (.58)	-.019 (1.35)	-.029 (2.12)	-.040 (2.35)	-.036 (1.44)
Per person family income (excluding wife's earnings)	.098 (2.42)	.078 (2.47)	.163 (4.32)	.064 (2.08)	.048 (1.78)	.049 (2.36)	.081 (3.82)	.051 (1.41)
Religion dummy	64.904 (2.14)	33.859 (.97)	52.926 (1.13)	162.431 (3.48)	148.127 (3.32)	58.960 (1.28)	-12.027 (.19)	38.595 (.57)
Language dummy	10.987 (.28)	71.796 (1.47)	-40.528 (.75)	-59.122 (.68)	-49.944 (.73)	-71.612 (.90)	107.960 (1.19)	208.111 (1.56)
Selection bias	409.289 (4.72)	329.504 (3.04)	552.956 (3.44)	323.988 (2.74)	378.497 (3.26)	496.736 (4.24)	376.411 (2.20)	322.514 (1.76)
Combined grouped $R^2 = .9591$[b]								
R^2	.03	.08	.05	.05	.07	.04	.04	.05
Standard error of regression	937	1091	1258	1329	1332	1294	1369	1471

Source: See Table 3
[a]See Table 3
[b]See Table 3

rather than provincial unemployment rates, or female rather than general unemployment rates, should have been used in our analyses.[9] As is the case for Gronau's results, but in marked contrast to Heckman's (1976, 1977) results, we find the selectivity bias in the offered wage equation to be significant with at least an 80 percent level of confidence for all age groups except 20–24 and 30–34.

One reassuring observation is that the sign of the coefficient of the bias correction term λ in the hours of work equation is always positive as required by our model. This was true too for the initial *OLS* estimates. Also, as required by our model, we find that the coefficients of the husband's earned income and of the family asset income are approximately equal for all age groups in the hours of work equation.

[9]Local unemployment rates for Canada by province and place of residence are not available, however. Nor would rates calculated from the Public Use Sample be suitable either, as these would be for the week prior to when the 1971 Census was taken rather than averages for 1970

The negative coefficients of the offered wage rate variable for the five younger age groups, all significant at at least an 80 percent confidence level, come as a surprise, however, since other researchers have found the response of the wife's hours of work to her wage rate to be positive.[10] Heckman (1976), in fact, estimates a model in which this response must be positive on theoretical grounds. He argues that, "Historical time-series and cross-section studies suggest that there is a monotonic *positive* relationship between wage rates and labor supply for married women ... so that excluding the 'backward bending' case is not objectionable ..." (1974, p. 681).[11] Yet looking at historical data for Canada[12] we find that, while both real wages for women and the female labor force participation rate have clearly risen over time, the percent of all female workers working forty weeks per year or more has fallen from 74.4 percent in 1961 to 67.0 percent in 1971 and the percent of all female workers working thirty-five hours per week or more has fallen from 90.0 percent in 1951 to 81.7 percent in 1961 to 71.3 percent in 1971.

In our model the response of the participation rate to a change in the offered wage is expected (though not constrained) to be positive. However, the sign of the offered wage variable in our hours equation cannot be determined a priori. Rather this coefficient is expected to reflect the balance between the positive substitution effects of increases in the offered wage rate on the number of hours of work, and a secondary negative effect which comes about because the more a woman is paid per hour the more rapidly the household budget constraint will be relaxed as she increases her hours of work. Thus our empirical findings that an increase in the offered wage rate will increase the probability of a wife working, but decrease her expected annual hours of work if she does work, are compatible both with the assumptions of our model and the observed historical developments for Canada. Moreover we will now argue that the empirical findings in other cross-sectional studies of a positive relationship between the wage rates of wives and their hours of work are the result of certain differences between these previous studies and ours.

As Hall (p. 151) points out, the overall or unconditional labor supply function can always be written as the product of a function for the probability of participating in the labor force (or of working in our study) and a conditional labor supply function defined for

[10]See, for instance, Hall, Heckman (1974, 1976), and Harvey Rosen.

[11]Heckman in his 1976 and 1977 papers estimates a reduced-form equation for the annual number of hours worked in which the offered wage rate does not explicitly appear. He then infers the sign of the impact of the offered wage on the annual number of hours worked from the signs of the coefficients of those variables which are included in his hours equation. He is able to make this inference because in moving from the theoretical appendix to the body of his 1974 paper, he drops without comment the offered wage, along with the price vector p, from his asking wage function. In a later paper (1978), Heckman explains that the asking wage in his one-period models is independent of the offered wage by assumption. Thus the coefficient of the *log* of the offered wage rate in his study becomes the constant of proportionality $1/\beta_5$ in our hours equation (14), which must be positive in both his model and ours since it represents the rate at which the annual hours of work adjusts to positive discrepancies between the offered wage rate and the asking wage rate at zero hours of work. In our model, however, the offered wage rate has not been dropped out of the asking wage function. Thus the coefficient of the *log* of the offered wage in our hours equation is $(1/\beta_5)\,(1 - \beta_4)$, as shown in Section. I.

[12]The numbers shown for the percentages of all female workers working forty weeks per year or more are calculated from the *1961 Census of Canada*, Volume II, Part 3, Table 22; and from the *1971 Census of Canada*, Public Use Sample-Individual File. The 1951 data on weeks worked are not comparable to the data for 1961 and 1971 since, in that year, part-time employment was converted to a full-time weekly basis. The data for both 1961 and 1971 are for wage and salary earners reporting the number of weeks worked in the previous year. The figures shown for the percentages of all female workers working thirty-five hours per week or more are calculated from the *1951 Census of Canada*, Volume V, Table 8; the *1961 Census of Canada*, Volume III, Part 3, Table 21; and the *1971 Census of Canada*, Volume III, Part 7, Table 32. The 1951 data are based on wage and salary earners 14 years of age and older who reported the number of hours worked in the week prior to enumeration. The data for 1961 and 1971 are based on wage and salary earners 15 years of age and older who reported the usual number of hours worked per week for the job held in the week prior to enumeration, or otherwise for the job of longest duration held since January 1 of the previous calendar year.

TABLE 6—ONE SET OF ESTIMATES FOR σ, β_5, AND β_4

	Age Groups							
	20–24	25–29	30–34	35–39	40–44	45–49	50–54	55–59
[illegible]	.720	1.217	1.177	1.059	.770	1.028	.974	1.010
[illegible]	.0030	.0044	.0015	.0027	.0011	.0013	.0007	.0024
β_4	1 709	2.718	1.304	1.627	1.434	.838	.724	.628

those who do participate (or do work in our study). We agree with Hall that for many purposes estimates of the overall labor supply function are required. We do not agree, however, that it is thus desirable to directly estimate the overall supply function.

From equations (10) and (12) in our model we see that the coefficient of the *log* of the offered wage in our function for the probability that a wife will work is $1/\sigma$, while the coefficient of this wage variable in our hours of work equation is $(1/\beta_5)(1-\beta_4)$. One set of indirect estimates for the parameters σ, β_5, and β_4 are shown in Table 6. The estimates of σ were obtained from the probit and offered wage coefficients of the education variable, while the estimates of β_5 were calculated using our estimated values for σ and the probit and hours equation coefficients for the employment income of the husband.[13] The resulting estimates for σ and β_5 are always positive, as required by our model. Moreover, despite the sign change from negative to positive between the age groups 40–44 and 45–49 for the estimated coefficients of *ln w* in our hours equations, our estimates for β_4 are always positive as hypothesized. Undue attention to the exact parametric specification of our model is unwarranted. However, the point is that if the coefficients of the wage rate variable in the function for the probability of working (or participating) and in the conditional labor supply function are not the same, then the direct estimation of an overall labor supply function will obscure the underlying behavioral responses to a change in the wage rate of the probability of working and the hours of work for those who do work.[14]

This point can best be made, perhaps, by jumping ahead to one of our examples given in Section VII. For a particular wife, we find that an exogenously induced increase in her estimated offered wage from $2.34 per hour to $3.28 increases her probability of working from 30.5 to 39.4 percent, but decreases her expected annual hours of work if she does work from 2,051 to 1,855 hours. However,

[13]To obtain the estimates of σ shown in Table 6, for each age group we divided our estimate shown in Table 4 of the appropriate element of the vector α_1 in equation (13) by our estimate shown in Table 3 of the appropriate element of the vector α_1/σ Then to obtain the estimates of β_5 shown in Table 6, for each age group we divided our estimate shown in Table 3 of $-\beta_3/\sigma$ in equation (12) by our estimate shown in Table 5 of $-\beta_3/\beta_5$ in equation (14) and then multiplied by the appropriate estimate of σ from Table 6. Finally to obtain the estimates of β_4 shown in Table 6, for each age group we multiplied our estimate shown in Table 5 of $(1-\beta_4/\beta_5)$ in equation (14) by the negative of the appropriate estimate of β_5 shown in Table 6 and added 1. The education and the employment income of the husband variables were used in estimating σ and β_5 because of the generally small standard errors and consistent behavior of the coefficients for these variables over age groups. It should be noted, however, that the coefficients of any variable appearing in both equations (12) and (13) could be used to obtain an estimate of σ, and any one of these estimates of σ could be used together with the coefficients of any variable appearing in both equations (12) and (14) to obtain an estimate of β_5.

[14]Most studies in which the overall labor supply function of wives is "directly" estimated, to use Hall's terminology, in fact involve a two-stage estimation procedure. First wage rates are imputed to wives who did not work, or to all wives in the sample, based on the observed wage rates of those wives with similar characteristics who did work. Then an hours of work equation is estimated for all wives in the sample, including those who worked zero hours (see, for instance, Hall and Rosen). The labor supply function estimated is thus conditional on the imputed, or imputed and observed, wage rates. Labor supply functions estimated in this manner are "overall," however, in the sense that they apply to all wives, not just those choosing to work. It should be noted too that the statistical problem which Hall cited as the major drawback of the estimation of a labor supply function conditional on the decision to work has been overcome in this paper, theoretically at least, by the introduction of a selection bias term as suggested by Heckman (1976).

simple calculation reveals that the *unconditional* expected number of hours of work for the wife in this example rises from 626 to 731 hours.

A second difference between our study and those of most other researchers lies in the choice of variables used to control for child status. In our own study we have included variables for the number of children younger than 6, the number of children 6–14 years of age, the product of the numbers of children in these two age groups, and the number of children ever born. On the other hand, Hall (1973) includes separate dummy variables for the presence of children younger than 7 only, the presence of children 7–13 only, and the presence of children in both age groups. And Heckman (1974, 1976, 1977) and Harvey Rosen only control for the number of children younger than 6. Moreover while we have analyzed the behavior of wives in five-year age groupings, Hall's study includes in the same analysis wives aged 20–59 and the studies by Heckman and Harvey Rosen include wives aged 30–44.

For our sample of working wives we find that the percentage who have children younger than 6 falls from 46.4 percent for wives 30–34 years of age, to 27.5 percent for wives 35–39 years old, to 12.1 percent for those aged 40–44. Moreover, of those working wives in these age groups who have children younger than 6, we find that 65.3, 81.1, and 79.4 percent, respectively, also have children 6–14 years of age. Yet neither Heckman nor Rosen differentiate between, say, a wife with one child younger than 6 and another wife with one child younger than 6 and three more children ranging from 6 to 16. Hall, on the other hand, fails to differentiate between wives with different numbers of children in the same age groups, and between older wives who have never had any children versus those whose children are all 14 years of age or older.

In Table 7 the working wives in our sample have been cross classified by their estimated offered wage rates and various measures of child status. For each cell containing thirty or more observations we show the mean annual number of hours worked and the mean number of children ever born.

When these wives are classified by their expected wage rates and the number of children younger than 6, as shown in columns 10 through 13 of Table 7, we find that there is a tendancy for wives 20–24 and 25–29 years of age who have higher estimated wage rates to work fewer, rather than more, hours. This tendancy is positive, however, for all of the older age groups. It should be noted moreover that after categorization of these wives by the number of children younger than 6 a strong negative association remains between the estimated wage rate and the number of children ever born, even for those wives in the two youngest age groups.[15] Nor is it intuitively surprising that wives with fewer children ever born also tend to have higher average numbers of hours worked than wives with more children ever born.

In columns 6 through 9 of Table 7, these same wives are cross classified by their estimated wage rates and their child status groups defined as 1) no children younger than 14, 2) children younger than 6 only, 3) children 6–14 only, and 4) children both younger than 6 and 6–14 years of age. Again we find a strong residual tendancy for wives with higher expected wage rates to have fewer children ever born. Thus our concurrent finding for this classification scheme that wives with higher estimated wage rates often work more hours than wives with lower estimated wage rates is, perhaps, to be expected.

Looking finally at the first five columns of Table 7, where these same working wives have been categorized by their estimated wage rates and their numbers of children ever born, we now find that wives younger than 45 with higher estimated wage rates tend systematically to work fewer hours. From Table 5 we see that these are the same five younger age groups for which the regression coeffi-

[15]In the exploratory phases of our analysis children ever born was used as an explanatory variable in our regression equations explaining the *log* of the offered wage rate. We replace this variable by the number of children younger than 6 because it was generally found to be insignificant in these regressions. Thus, the strong negative relationship between the estimated wage rate and the number of children ever born is believed to be due to the common influence of some third variable such as education.

TABLE 7—MEAN HOURS WORKED AND CHILDREN EVER BORN FOR WORKING WIVES CLASSIFIED BY THEIR ESTIMATED WAGE RATES AND VARIOUS MEASURES OF CHILD STATUS

Estimated Wage Rate	Number of Children Ever Born					Child Status Group				Number of Children <6			
	0	1	2	3	4+	None <14	<6 Only	6-14 Only	<6 and 6-14	0	1	2	3+
20-24													
<$2.50	1576[a] 0.0[b] (1353)[c]	1125 1.0 (800)	920 2.0 (253)	810 3.0 (49)		1583 0.0 (1356)	1060 1.2 (1046)		840 2.1 (41)	1578 0.1 (1376)	1099 1.0 (852)	920 2.0 (210)	
≥$2.50	1446 0.0 (370)					1445 0.0 (373)				1445 0.0 (373)			
25-29													
<$2.50	1587[d] 0.0 (232)	1214 1.0 (283)	956 2.0 (437)	952 3.0 (185)	843 4.6 (81)	1681 0.2 (219)	994 1.5 (467)	1276 2.1 (177)	888 2.7 (355)	1500 1.0 (396)	1061 1.7 (486)	788 2.4 (288)	775 3.6 (48)
≥$2.50	1663 0.0 (722)	1115 1.0 (446)	881 2.0 (223)	841 3.0 (36)		1667 0.0 (724)	966 1.2 (543)	1455 1.4 (72)	989 2.3 (94)	1648 0.2 (796)	1044 1.2 (478)	764 2.1 (154)	
30-34													
<$2.50	1574 0.0 (125)	1311 1.0 (159)	1134 2.0 (360)	1068 3.0 (282)	1048 4.7 (238)	1543 0.3 (129)	1013 1.4 (104)	1250 2.5 (563)	967 3.4 (368)	1304 2.1 (697)	975 2.7 (360)	969 3.7 (98)	
≥$2.50	1615 0.0 (158)	1203 1.0 (177)	997 2.0 (223)	947 3.0 (134)	863 4.4 (43)	1640 0.0 (150)	977 1.4 (203)	1313 2.0 (175)	870 2.7 (207)	1464 1.1 (325)	1008 1.9 (293)	704 2.4 (102)	
35-39													
<$2.50	1569 0.0 (106)	1320 1.0 (119)	1159 2.0 (329)	1156 3.0 (308)	1013 5.1 (409)	1540 0.9 (161)	814 1.7 (52)	1197 3.1 (754)	928 4.1 (304)	1258 2.7 (915)	948 3.6 (291)	684 4.2 (54)	
≥$2.50	1389 0.0 (83)	1310 1.0 (73)	1133 2.0 (124)	989 3.0 (110)	1160 4.6 (94)	1538 0.4 (91)	892 1.3 (39)	1180 2.7 (267)	913 3.3 (87)	1271 2.1 (358)	915 2.5 (90)	916 3.0 (33)	
40-44													
<$2.50	1500 0.0 (149)	1395 1.0 (164)	1288 2.0 (350)	1243 3.0 (306)	1125 5.1 (518)	1502 1.7 (481)	1074 2.8 (41)	1169 3.4 (805)	991 4.7 (160)	1294 2.8 (1286)	985 4.2 (169)		
≥$2.50	1425 0.0 (42)	1304 1.0 (43)	1128 2.0 (102)	1004 3.0 (101)	1159 4.7 (121)	1400 1.6 (107)		1069 3.1 (255)	1137 4.0 (37)	1167 2.6 (362)	1078 3.5 (45)		
45-49													
<$2.50	1464 0.0 (139)	1452 1.0 (155)	1313 2.0 (327)	1354 3.0 (266)	1261 5.2 (410)	1421 2.1 (768)		1255 3.9 (465)	960 5.4 (46)	1358 2.8 (1233)	909 4.8 (59)		
≥$2.50	1486 0.0 (97)	1486 1.0 (71)	1238 2.0 (165)	1313 3.0 (124)	1093 4.9 (146)	1401 1.7 (360)		1126 3.5 (239)		1291 2.4 (599)			
50-54													
<$2.50	1484 0.0 (138)	1302 1.0 (139)	1296 2.0 (270)	1243 3.0 (198)	1232 5.3 (306)	1324 2.4 (816)		1188 4.2 (219)		1295 2.7 (1035)			
≥$2.50	1716 0.0 (55)	1681 1.0 (46)	1300 2.0 (71)	1153 3.0 (49)	1264 5.0 (80)	1475 1.9 (225)		1186 4.0 (74)		1403 2.4 (299)			
55-59													
<$2.50	1467 0.0 (83)	1263 1.0 (91)	1262 2.0 (166)	1186 3.0 (129)	1259 5.4 (212)	1274 2.7 (635)		1234 5.0 (46)		1272 2.9 (681)			
≥$2.50	1387 0.0 (61)	1390 1.0 (48)	1357 2.0 (79)	1344 3.0 (51)	1253 4.6 (45)	1363 1.9 (271)				1350 2.0 (284)			

[a]Mean annual number of hours worked.
[b]Mean number of children ever born.
[c]Number of observations.
[d]*Underlining is used to denote categories where the mean number of hours worked is found to increase with increases in the estimated wage rate.*

cients of our wage rate variable were found to be significantly negative. Table 7 also reveals that the exceptions to this pattern for wives younger than 45 all occur in the child status categories of 0 and 4+ children ever born. These two categories deserve further comment.

In the category for 4+ children ever born, there is still a persistent tendancy for wives with higher estimated wage rates to have smaller mean numbers of children ever born than wives with lower estimated wage rates. Thus again the finding that wives in this child status category tend to work more hours if their estimated wage rates are higher is not surprising.

In the case of wives with no children ever born, the problem is different. Looking at the row in Table 7 for wives 25–29 who earn less than $2.50, we find there are 396 wives with no children younger than 6, 219 wives with no children younger than 14, and 232 wives with no children ever born. The number of children ever born to a women includes all live births, even if the children born have since died, but excludes children who were born to other women but who are living with the wife in question. Thus we find that of the 232 zero parity wives referred to above, 199 have no children younger than 14 living with them, but 25 have children younger than 6 living with them, 5 have children aged 6–14 living with them, and 3 have both children younger than 6 and children aged 6–14 living with them. Moreover the corresponding mean numbers of hours of work for these different categories of zero parity wives are 1,651, 1,166, 1,341, and 1,224. This example underlines the extreme difficulty of adequately controlling for child status by the use of any single variable.

One further distinguishing feature of our study is that we have not included education in our asking wage equation, and hence it does not appear as a separate variable in our hours equation. This variable was dropped from our asking wage equation largely because the inclusion of both this variable and the estimated offered wage in our hours equation resulted in severe problems of multicollinearity.

Whatever the biases may be in our estimates of the impact of the offered wage on hours of work, it is interesting to note that the associated wage elasticities evaluated at the mean hours of work for the age groups 20–24 through 55–59 are −.194, −.313, −.173, −.199, −.320, .094, .299, and .120, respectively. While all of our estimates are well below the range of positive elasticities reported by other researchers for married women, they seem broadly consistent with Orley Ashenfelter and Heckman's estimate of −.15, Sherwin Rosen's estimates of −.30 to −.07, Finegan's estimates of −.35 to −.25, and John Owen's estimates of −.24 to −.11, where all of these estimates are for men.

VI. Explanatory Power of the Estimated Model

Looking at the pseudo R^2s and the R^2s shown in Tables 3–5, our estimated equations seem to explain very little about the labor force behavior of individual married women. This is to be expected since we do not have data on many of the factors which have important effects on the lifetime career patterns of individual women. However, the variables included in our estimated relationships should capture some of the *average* or macro differences between groups of wives.

To check the extent to which this is true, we first used the relationships shown in Table 3 to predict whether or not each married woman in our data sample worked in 1970. We then grouped these women according to age (15–24, 25–34, 35–44, 45–64, 65+), education (< 12 years, complete high school, bachelor or first professional degree), presence of children in different age groups (none < 14, < 6 only, 6–14 only, < 6 and 6–14), earned income of husband plus family asset income (< $3,000, $3,000–$5,999, $6,000–$8,999, $9,000–$11,999, $12,000–$14,999, $15,000+), place of residence (urban, rural), and region (Maritimes, Quebec, Ontario, Prairies, British Columbia). For each of the resulting 3,600 groups we computed the proportion of wives predicted to work and the proportion who actually worked, and regressed the predicted on the actual proportions of working wives using *GLS*

because the 3,600 groups do not all contain the same number of observations. The combined grouped R^2 calculated in this manner is .9072. Likewise we used the relationships shown in Table 4 to calculate an offered wage for each of the wives in our data sample who worked in 1970, and computed the mean actual and estimated wage rates for the wives in each of our 2,880 groups. (The age categories are now 20–24, 25–34, 35–44, and 45–59.) The R^2 for the *GLS* regression of the estimated on the actual mean wage is 7307. Using the estimated wage rates already calculated and the relationships shown in Table 5, we now calculated a combined grouped R^2 of .9591 for the annual number of hours worked, and a combined grouped R^2 of .6254 for the estimated annual income.

VII. Policy Conclusions

Our estimation results indicate that variables under the control of individual wives, or these wives and their families, do have substantial impacts on the labor force behavior and earned incomes of these wives. It is difficult, however, to jump from the presentation of empirical results in the preceding sections to a discussion of the implications of these results for public policy. In the first place, large-scale government programs would have substantial secondary impacts through changes in the macro economy which cannot be explored using a model which treats macro phenoemena as exogenous. Secondly the causal significance of our findings is often unclear. For instance, are large numbers of women with small children not working because of the associated "costs" of working as we have postulated, or are they not working because of tastes and preferences which lead them to have both high asking wages and the desire to care for their own children. Government programs also entail costs, and no cost information is presented in this paper. Moreover the impacts of a government program would not generally be limited simply to changes in the propensity to work, wage rates, hours of work, and incomes. Concepts like self-fulfillment, role models, the right of women to have control over their own bodies, equal pay for equal work, and so forth really lie outside the scope of this analysis. However, with these qualifications firmly in mind, certain tentative observations can still be made.

We will make these observations in the context of a numerical example. Consider an average 41-year-old wife living in a small New Brunswick city with three children aged 16, 13, and 4, and a husband whose earned income is $8,400. Assume that the asset income of the family is $373 per year, the wife was 21 years old when she got married, and she completed nine years of formal education. Our estimated relationships predict that this wife has a 30.5 percent probability of working, and that if she does work she will earn $2.34 per hour, work 2,051 hours per year, and have an annual income of $4,810.

Suppose now that this wife had stayed in school three more years, or had completed three more years of schooling in a continuing education program for adults. Our equations predict that with a grade 12 rather than a grade 9 education she would have a probability of working of 37.1 percent, an hourly wage of $3.01, a work year of 1,902 hours, and an annual income of $5,735. Or suppose that the birth of the third child had been prevented through improved usage of contraceptives, a vasectomy or tubal ligation, or an abortion. Now the wife's expected probability of working would be 43.2 percent, her expected hourly wage rate would be $2.83, her expected work year would be 2,150 hours, and her expected annual income would be $6,087. Alternatively, suppose that this family moved to a city with a population over 30 thousand in Ontario, with the husband's income remaining unchanged. The 1970 unemployment rate for Ontario is 4.3 percent compared with 8.0 percent for New Brunswick. Also the 1970 value of our local opportunity for jobs index for large cities in Ontario is .49 compared with .39 for New Brunswick. In this more favorable labor market the wife's expected probability of working, wage rate, work year, and annual income are 42.9 percent, $2.83, 1,889 hours, and $5,346, respectively.

For wives who cannot move, of course,

local labor market conditions are givens which can only be altered through direct government spending, legislation, court rulings, or economic growth. Suppose the coefficients of the variable for the number of children younger than 6 in the asking wage function were reduced in size by 20 percent through government subsidies to day care, tax credits for child care expenses, or reduced regulation of private day care. Now our New Brunswick wife would be expected to have a 33.4 percent probability of working, a wage rate of $2.47, a work year of 2,077 hours, and an annual income if she works of $5,129. Or suppose that the constant term in our function for the *log* of the offered wage were exogenously increased by .1823 through equal pay legislation, raising the mean wage for wives aged 40–44 fairly close to the mean wage for men in this age group. Now we would expect this same wife to have a probability of 39.4 percent of working, a wage rate of $3.28, a work year of 1,855 hours, and an annual income of $6,079.[16] Another legislative or judicial happenstance might be quotas for women by occupation. To take an extreme example, suppose it were required that 50 percent of the jobs in all occupations be held by women, and that this could be accomplished without any change in the unemployment rate or the total number of workers in each occupation. Then the value of our opportunity for jobs index would rise from .39 to .54. If the employment income of her husband and the family asset income remained unchanged as well, we would now expect our New Brunswick wife to have a probability of 42.1 percent of working, a wage rate of $2.97, a work year of 1,875 hours, and an annual income of $5,574.

Economic growth has been largely ignored as a factor affecting the labor force behavior and earned incomes of women. Our results indicate, however, that if the total number of jobs available in small New Brunswick cities in the Clerical, Sales, and Service occupations were for instance doubled, the value of the local opportunity for jobs index would rise from .39 to .60. As a result we would expect our 41-year-old wife to have a probability of working of 47.2 percent, a wage rate of $3.27, a work year of only 1,805 hours, and an annual income of $5,900. Thus substantial improvement in both the probability of working and the accompanying monetary rewards can be obtained within the existing structure of occupational segregation through economic growth. Not all types of growth will produce these results, however. If this same unrealistically large number of new jobs were created in Construction rather than the Clerical, Sales, and Services occupations, the value of our local opportunity for jobs index would still be .39 to two significant places, and the expected labor force behavior and potential annual income of our New Brunswick wife would remain virtually unchanged.

Finally, economic conditions beyond the control of a wife or her family may also affect her labor force behavior through their impact on the earned income of her husband. Suppose that the husband of our hypothetical wife earned only $6,400 instead of $8,400 Now we would expect this wife to have an increased probability of working of 32.6 percent, a wage rate of $2.44, an increased work year of 2,094 hours, and an annual income of $5,110.

The impacts of these various exogenous changes on the expected wage rate, labor supply, and earned income of our hypothetical New Brunswick wife are summarized in Table 8. One obvious observation to be made from the figures presented in Table 8 is that all of the changes considered, except an increase in construction jobs, do increase the probability of working for wives like the one in our example. These same changes also all lead to increases in the unconditional expected labor supply of these wives, which is the product of the estimated probability of working and the expected annual hours of work given the decision to work. The welfare, or utility, implications for these wives of these various changes may be quite different, however. For instance, few wives would see a fall in their husband's income as improving their welfare. In fact, government programs designed to improve the job prospects and wage rates of low-income men may well be the first choice of wives working relatively

[16]The appropriate estimate of σ shown in Table 6 was required to compute these estimates

TABLE 8—EXPECTED IMPACTS OF VARIOUS EXOGENOUS CHANGES ON THE WAGE RATE AND LABOR SUPPLY OF A HYPOTHETICAL WIFE[a]

Changes[b]	Expected Probability of Working (percent)	Expected Offered Wage (dollars)	Conditional Expected Annual Hours of Work	Conditional Expected Income of Wife (dollars)	Unconditional Expected Annual Hours of Work
Control	30.5	2.34	2,051	4,810	626
Three more years of education	37.1	3.01	1,902	5,735	706
Birth of third child prevented	43 2	2.83	2,150	6,087	929
Moved to large city in Ontario	42.9	2.83	1,889	5,346	810
Subsidized day care for third child	33.4	2.47	2,077	5,129	694
Drastic equal pay legislation	39 4	3 28	1,855	6,079	731
Drastic occupational quotas for women	42.1	2.97	1,875	5,574	789
Drastic increase in clerical, sales and service jobs	47.2	3.27	1,805	5,900	852
Drastic increase in construction jobs	30.5	2 34	2,051	4,810	626
Husband's income reduced to $6,400	32.6	2 44	2,094	5,110	683

[a]This wife is 41-years old, has nine years of education, married at age 21, has a husband who earns $8,400, and lives in a small city in New Brunswick in a family with three children aged 4, 13, and 16, and $373 of asset income
[b]See text for details.

long hours to compensate for the low incomes of their husbands.

Child care and birth control programs primarily benefit those wives who take advantage of the child care subsidies or who would otherwise have unwanted pregnancies. Older wives; infertile wives; and wives who prefer informal to paid child care, or who already have access to all the medical help they want in controlling their own fertility, will not be affected by these programs. Moreover the welfare of the individual wives using these services might well be higher if the money spent on these programs were instead distributed among them as aid payments which could be spent according to the needs of each family. The justification for public funding of programs of this type must lie then in the efficiency gains or in the personal and social externalities associated with the public provisions of certain tied benefits and services.

The remaining types of changes for which figures are shown in Table 8 all serve to raise the offered wage, and would presumably benefit wives of all ages and child statuses. The resulting increases in earning power could be spent as desired by each family. Moreover, measures which raise the offered wage, as opposed to lowering the asking wage, allow wives the choice of increasing or maintaining their levels of earnings while spending more time with their families or on other nonmarket activities.

Strictly on the grounds of individual welfare, therefore, it would appear that government policies designed to raise the offered wage rates of women are to be preferred to programs such as subsidized child care which primarily affect the asking wage. Within these two categories of measures there are also differences in the magnitudes of the expected labor force responses. The prevention of the birth of an unwanted child, for instance, would appear to result in a higher level of economic well-being for the family involved than the subsequent subsidization of care for this child following its birth. For desired children, of course, this tradeoff does not exist.

Likewise the above figures suggest that for many individual wives it would require the imposition of drastic equal pay legislation or job quotas, or massive creation of jobs in female-dominated occupations, to yield the same increase in offered wage rates as a few more years of education. This suggests that a higher priority should be given to the provision of free continuing adult education, and the availability of scholarships and loans for women who wish to attend postsecondary

educational programs either immediately following graduation from high school or later in their lives. Many wives might like to return to school after their youngest child has entered primary school, for instance. Of course, if large numbers of women obtained additional education this would eventually serve to depress the returns to this education.

One final observation is that stimulation of female-dominated industries or occupations may have many of the same desirable effects on the employment prospects for women that more equal pay legislation or job quotas by sex would have, without the uncomfortable accompaniments of interfering through the legislative and judicial systems in individual hiring decisions and negotiations. In this respect, the economic development of the last few decades has been a strong supporter, or perhaps even an instigator, of the women's movement.

REFERENCES

O. A. Ashenfelter and J. J. Heckman, "Estimating Labor Supply Functions," in Glen G. Cain and Harold W. Watts, ed., *Income Maintenance and Labor Supply*, Chicago 1973, 323–52.

B. R. Bergmann and I. Adelman, "The Economic Role of Women," *Amer. Econ. Rev.*, Sept. 1973, *63*, 509–14.

William G. Bowen and T. Aldrich Finegan, *The Economics of the Labor Force Participation*, Princeton 1969.

Abbott L. Ferris, *Indicators of Trends in the Status of American Women*, New York 1971.

T. A. Finegan, "Hours of Work in the United States: A Cross-Sectional Analysis," *J. Polit. Econ.*, Oct. 1962, *70*, 452–70.

V. R. Fuchs, "Recent Tends and Long-Run Prospects for Female Earnings," *Amer. Econ. Rev. Proc.*, May 1974, *64*, 236–42.

I. Garfinkel, "On Estimating the Labor-Supply Effects of a Negative Income Tax," in Glen G. Cain and Harold W. Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973, 205–64.

R. Gronau, "The Intrafamily Allocation of Time: The Value of the Housewives' Time," *Amer. Econ. Rev.*, Sept. 1973, *63*, 634–51.

———, "Wage Comparison—A Selectivity Bias," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1119–43.

E. Gross, "Plus a Change . . . ? The Sexual Structure of Occupations Over Time," *Soc Probl.*, Fall 1968, *16*, 198–208.

M. Gunderson, "Work Patterns," in Gail C. A. Cook, ed., *Opportunity for Choice: A Goal for Women in Canada*, Ottawa 1976, 93-142.

R. E. Hall, "Wages, Income and Hours of Work in the U.S. Labor Force," in Glen G. Cain and Harold W. Watts, ed., *Income Maintenance and Labor Supply*, Chicago 1973, 102–62.

J. J. Heckman, "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, July 1974, *42*, 679–94.

———, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals Econ. Soc. Measure.*, Fall 1976, *5*, 475–92.

———, "A Partial Survey of Recent Research on the Labor Supply of Women," *Amer. Econ. Rev. Proc.*, May 1978, *68*, 200–07.

———, "Sample Selection Bias as a Specification Error with an Application to the Estimation of Labor Supply Functions," work. paper no. 172, Nat. Bur. Econ. Res., Mar. 1977.

H. G. Lewis, "Comments on Selectivity Biases in Wage Comparisons," *J. Polit. Econ.*, Nov./Dec. 1974, *82*, 1145–55.

R. T. Michael, "Education in Nonmarket Production," *J. Polit. Econ.*, Mar./Apr. 1973, *81*, 306–27.

A. Nakamura, M. Nakamura, and D. Cullen, *Employment and Earnings of Married Females*, Ottawa 1979.

Valerie K. Oppenheimer, *The Female Labor Force in the United States: Demographic and Economic Factors Governing its Growth and Changing Composition*, Berkeley 1970.

Sylvia Ostry, *The Occupational Composition*

of the *Canadian Labor Force*, Ottawa 1967.

J. D. **Owen**, "The Demand for Leisure," *J. Polit. Econ.*, Jan./Feb. 1971, *79*, 56–76.

H. S. **Rosen**, "Taxes in a Labor Supply Model with Joint Wage-Hours Determination," *Econometrica*, May 1976, *44*, 485–507.

S. **Rosen**, "On the Interindustry Wage and Hours Structure," *J. Polit. Econ.*, Mar./Apr. 1969, *77*, 249–73.

Statistics Canada, *Ninth Census of Canada, 1951*, Vols. IV, V, Ottawa 1953.

———, *1961 Census of Canada*, Vol. II, Part 3; Vol. III, Part 1, Ottawa 1963.

———, *1971 Census of Canada*, Vol. III, Part 2, Ottawa 1976.

Public Employee Market Power and the Level of Government Spending

By Paul N. Courant, Edward M. Gramlich, and Daniel L. Rubinfeld*

Recent budgetary rhetoric emanating from Washington and other governmental capitals suggests a growing fear that public spending is getting out of control. For long periods of time the government budget has grown more rapidly than *GNP* in most mixed economies, and observers of these trends have begun to realize that if this process continues, public expenditures will approach very high shares of *GNP* and income tax rates could get close to unity.[1]

These scare stories are counteracted by the simple question that if government gets too large, why can't voters band together to stop its growth? Rational, informed, democratic voting processes should provide a limit to the size of the public sector; indeed they should insure that the public sector is just as large as the voters want it to be. According to what economists have come to know as the "median voter" theory, it is puzzling to know exactly how government spending could ever get too high or out of control.

There have been several attempts to explain the apparent anomaly. The major focus of previous efforts has been on some aspect of bureaucratic aggrandizement, either broadly or narrowly construed. William Niskanen (1971), for example, presents a model in which bureaucracies desire to obtain as large a budget as possible for the bureau in which they are employed. (See also his 1975 paper.) Despite competition from other bureaus, the size of the overall governmental budget is larger than socially optimal because the nature of the budget process allows bureaus to act as price-discriminating revenue maximizers. Their ability to use their market power is constrained, both by competition from other bureaus and by the preferences of relevant legislative committees. As is implicit in the title of his work, *Bureaucracy and Representative Government*, Niskanen's major concern is with the way in which the institutions of representative government (particularly the *U.S.* federal government) may lead to an overprovision of public services. The model is not directly relevant to the behavior of local governments since it ignores two important constraints on local government spending. One is provided by households' opportunity to vote directly on referenda concerning tax collections, and the other by the ability of households to leave local jurisdictions in response to expenditure-taxation packages which they find to be unsatisfactory.[2]

More general in application than Niskanen's work are a number of papers which focus on the ability of public employees to influence the political process so as to increase both wages and the size of the public sector.[3] The implications of this approach have been discussed by a number of authors, but in each case the underlying model has been left unstated or undeveloped. For example, James Buchanan considers the possible ramifications of the right of public employees to vote when he argues:

*Department of economics and Institute of Public Policy Studies, University of Michigan. We would like to thank Theodore Bergstrom, George Borts, Harvey Brazer, Alan Deardorff, Arthur Okun, A. Mitchell Polinsky, and a referee for helpful suggestions.

[1]Much of this rhetoric is contained in popular articles: see Norman MacRae and Jude Wanniski The relevant formal empirical literature involves a discussion of Adolph Wagner's law by such authors as Richard Wagner and Warren Weber and Richard Bird.

[2]A related approach, emphasized by political scientists but rarely by economists, concentrates on the voting pressure of the clientele groups of government programs One economist's treatment of this issue is by Richard Craswell

[3]The idea was treated in passing by Melvin Reder. See also John Pencavel. The empirical work of such authors as Ronald Ehrenberg and Gerald Goldstein, and Sharon Smith is also indirectly relevant.

> Bureaucrats are no different from other persons, and, like others, they will rationally vote to further their own interests as producers when given the opportunity. Clearly their interests lie in an expanding governmental sector, and especially in one that expands the number of its employees. Salaries can be increased much more rapidly in an expanding agency than in a declining or stagnant one. [p. 14]

Buchanan clearly implies that bureaucratic size and market power are highly correlated, but doesn't consider any limits to the growth of government. The dynamics of bureaucratic and governmental growth are also analyzed by Gordon Tullock, who argues that bureaucrats will utilize their market power first to expand the size of the bureaucracy and then to increase wages. Once again, no explicit model is provided to permit consideration of forces that might check the growth of government.

A number of authors have extended the earlier work of Buchanan, Tullock, Niskanen, and others by focusing more specifically on behavioral and motivational differences between public and private employees. For example, Winston Bush and Arthur Denzau cite evidence that voter participation (and presumably support for the public sector) is higher for bureaucrats than for private sector voters, and conclude that higher public sector growth may be the result. A related argument is made in the paper by Thomas Borcherding, Bush, and Robert Spann, who suggest that bureaucrats view public goods as yielding higher (wage) income as well as utility from consumption. If public employees perceive this added dimension as a reduction in the price of public goods, then public employees will opt for a larger public sector.

While certainly not complete, this brief overview of the literature is suggestive. A number of reasons have been given as to why public sector growth might get out of control, without much discussion of the possible checks on governmental growth. As Buchanan and Tullock themselves state, "Presumably there is some limit on this process, but it has not been determined either theoretically or empirically" (p. 150).

In this paper we attempt to respond to the theoretical gap recognized by Buchanan and Tullock by providing a model that permits explicit analysis of some of the issues just raised. The context of the model is a local government beset by growing political and economic power of its own employees on one side, and the threat of mobility of the private sector on the other. As regards the former, it is natural, but not necessary, to view the process in terms of a cohesive public employee union that can bargain for uniform (and high) public wages, and also can choose to vote for a larger public employee work force (though we will see that this particular behavior can be suboptimal). In terms of the classic seller-buyer dichotomy that underlies almost all of economics, to the extent that public employees or their unions gain political power over the budgetary behavior of the jurisdiction, the problem becomes interesting because the suppliers of public goods are in part their own demanders, with the private sector having little to do but pay the bills.

Section I presents the assumptions of the model. Of particular importance is the assumption that public employees have some control over both their own wages and the level of public output. Section II shows how the size of the government budget and its composition between wage rates and employment are determined, first when the private sector workers are the dominant electoral bloc, and then when public sector workers assume control. The main object of interest in this latter case is the ratio of government spending to total income of the community, which will be shown to depend ultimately on the sensitivity of private sector location decisions to the cost of government services. In Section III we deal explicitly with how the competing demands of public and private sector employees might be resolved by a majority rule voting process, noting that public employee bargaining power can alter the level of public employment and public expenditures even when these employees are a minority of the total voting population. This has important implications for the median voter theory, at least in its simplified form, because it argues that even when the median voter is a private employee, the presence of

public sector market power will result in a level of public sector expenditures which is influenced by public sector voting power as well as bargaining power. At the same time, the influence of public employees should be a good deal less than is often claimed in the popular press, because the threat of outmigration by private employees constrains the size of the public budget. In addition, for a given public budget, as the employee work force gets large, public wages must fall. Finally, in Section IV, we conclude by summarizing the implications of the analysis for the question of the controllability of public budgets and suggest some further policy issues that might be examined.

I. Assumptions of the Model

To clarify the analysis which follows, we first list the assumptions used in the paper, grouping them by topic.

ASSUMPTION 1. *Employee Optimization:* All actors in the model are employed in either the public or private sector,[4] and maximize utility functions with the arguments private consumption (C) and public output (E_g). The latter is measured exactly by the level of public employment. This assumption is the basis for the voting and private spending behavior of all employees.

ASSUMPTION 2. *Mobility and Tastes:* We assume that all private sector employees have identical tastes for public and private output, and all public sector employees have identical tastes, though the tastes of the public and private employees will in general not be the same.[5] Since a major theme of the paper is the effect of private sector mobility on the size of the government budget, we assume that although private employees have identical tastes for output, their underlying desires to live in the community vary randomly. When things become sufficiently disadvantageous, those members of the private work force with the weakest desires to live in the community will leave.[6]

ASSUMPTION 3. *Goods and Labor Markets:* We make what might be known as small-country assumptions regarding the behavior of goods and labor markets. Private sector workers produce consumption goods (C) which are sold on a national market at a fixed price which we normalize at unity. All returns from the sale of these consumption goods are distributed equally to the private employees (E_p). Hence the private wage bill W_p (the private money wage) times E_p equals the gross value of private output (C).[7] Over the relevant range, there are no diminishing returns to private sector labor, so W_p remains constant. All private sector workers are assumed to supply labor inelastically as long as they remain in the community.[8]

ASSUMPTION 4. *The Public Sector:* We assume that the government must always balance its budget. Total revenues, the product of a tax rate and the tax base, must equal total expenditures, the product of the number of government employees (E_g) and their

[4]Implicit in this assumption is the existence of some mechanism (such as a union) whereby public employees determine who will work in the public sector It is further assumed that once in the public sector, all public employees get paid the same wage, which, due to the limitation on entry, could exceed the private sector wage. The model could be extended to deal with seniority rules of various kinds.

[5]If an individual changes sectors, he is assumed to acquire the tastes associated with his current sector, either because of a socialization process or because he now has the same interests as others in his current sector. The assumption of identical tastes within sectors will be relaxed in Section IV, where we consider distributions of tastes. Even there, however, we retain the assumption that if an individual changes sectors his tastes are drawn from the distribution associated with his current sector

[6]We do not consider the potential mobility of public employees because our major interest is in private sector responses to the exercise of public sector market power Whatever public employees are likely to do in response to the exercise of such power, it is unlikely that they will leave the community and thus abandon the fruits of their ability to exploit the private sector.

[7]Nonwage income can be incorporated into the analysis by viewing the private wage W_p as the gross pretax income of private workers from all sources, including capital ownership and returns to entrepreneurial services.

[8]It would be possible to incorporate a variable labor supply, but the qualitative results will differ little from those when private workers are free to leave the community.

annual money wage rate (W_g). The nominal tax base is simply total community wage income (both private and public). The tax rate on income is assumed to be proportional and uniform across all employees, though the model could easily be adapted to cases of progressive or regressive taxes. The level of government output (E_g) is determined by the majority voting of utility-maximizing citizens.

ASSUMPTION 5. *Income and the Tax Base*: Since all income in the community is wage income, total income may be represented by an index of wages (W^*) times total employment ($E = E_g + E_p$).[9] Thus,

$$W^*E \equiv W_pE_p + W_gE_g \tag{1}$$

Note that when government employees have no bargaining power, $W^* = W_g = W_p$. With bargaining power, W_g can exceed W_p, and W^* becomes a weighted average of wage rates in the two sectors.

In the competitive case where employees have no bargaining power, all wages are taken as given and the tax base is fixed and independent of the composition of output between the public and private sector. When employees have bargaining power, upward changes in W_g will alter W^*E for two offsetting reasons: a) since private wage rates are fixed by the price of consumer goods, increases in W_g at a given E_g will tend to raise W^*E through the W_gE_g term; and b) since private employees are allowed to leave the community in response to monopolistic behavior by public employees, declines in E_p at a given W_p will lower W_pE_p and thus W^*E.

II. The Model

The results of the model are described by first calculating outcomes when private employees are the dominant voting bloc—giving the standard outcome of prevailing public finance theories. We then contrast these results to those that are obtained when public employees are given control over public output, public wages, or both together.

A. *Private Employees*

Private sector employees are assumed to maximize a utility function of the form

$$U_p = U_p(E_g, C_p) \tag{2}$$

where U_p is the level of utility achieved by private employees when optimizing with respect to the level of public sector employment and private consumption of private employees (C_p). The utility function is maximized subject to the budget constraint

$$W_p = C_p + P_pE_g \tag{3}$$

where P_p is the price that the private employee must pay for the hiring of an additional public servant, and the private employee is assumed to be a price taker with respect to both private consumption goods and public employment.

The price paid by a private sector employee for a unit of public output is equal to the product of the cost of a public employee (the public sector wage, W_g) times the *share* of community-wide taxes paid by the private sector voter. If taxes are assessed on a per capita basis for all N residents of the jurisdiction, this price (W_g/N) is independent of the income of the private sector employee. However, with our assumption of a proportional income tax,[10] the share of wages the private sector voter must pay is W_p/W^*E, where W^*E is the tax base of the community.[11] The price of a public employee is thus given by

$$P_p = W_gW_p/W^*E \tag{4}$$

[9] In effect, we have assumed that both sectors produce output by means of a constant returns-to-scale one-factor technology. Explicit consideration of two-factor production functions in either one or both sectors would both complicate and enrich the model. For example, one could examine a situation in which factor intensities differed in the two sectors, in which case changes in the allocation of labor would induce changes in the private sector wage. While such possibilities could be important, for the sake of brevity we do not treat them here.

[10] This approximates Joseph Pechman and Benjamin Okner's findings for the *U.S.* local government sector.

[11] If wage rates are fixed in the public sector but not the same in both sectors, changes in E_g voted on by private employees will reallocate labor between the two sectors and make a slight alteration in the nominal tax base of the community and therefore in the tax price for private employees. We assume that private employees are unaware of this effect and simply take W^*E as given.

After substituting (4) into (3), we obtain the budget constraint for the private employee:

$$C_p = W_p\left(1 - \frac{W_g E_g}{W^*E}\right) \tag{5}$$

The private employee assumes that the tax base W^*E and the prices he faces are not affected by his actions and maximizes (2) subject to (5). Solving the first-order condition yields

$$\frac{U_1}{U_2} = \frac{W_g W_p}{W^*E} = P_p \tag{6}$$

This is the standard result (appropriate for both voters in political elections and private consumers) that the individual will desire to consume that amount of public goods which equates the marginal rate of substitution with the price ratio.

In the case of the CES utility function, for example, the utility function is

$$U_p = [aE_g{}^r + (1-a)C_p{}^r]^{1/r} \tag{7}$$

and the optimal level of public employment of condition (6) is

$$E_g = \left\{\left[\frac{a}{1-a}\right]^{1/(1+r)}\left[\frac{1}{W_p}\right]^{r/(1+r)} \cdot \left[\frac{W_g}{W^*E}\right]^{1/(1+r)} + \frac{W_g}{W^*E}\right\}^{-1} \tag{8}$$

where $r = (1-s)/s$, s being the elasticity of substitution, and a is the distribution parameter. It is easy to show that E_g is homogeneous of degree zero in all prices. In addition, $\partial E_g/\partial W_g$ will always be negative. The desired level of public employment arising from equations (6) and (8) is equivalent to the level of public output resulting from the familiar median voter model in which the median voter is a private employee. (See, for example, James Barr and Otto Davis; Theodore Bergstrom and Robert Goodman; Borcherding and Robert Deacon.)

B. *Public Employees*

The constrained maximization exercise for public employees is very similar. Employees are assumed to maximize

$$U_g = U_g(E_g, C_g) \tag{9}$$

subject to

$$W_g = C_g + P_g E_g \tag{10}$$

where C_g is the private consumption of the public employees and P_g is the tax price facing public employees; i.e.,

$$P_g = W_g^2/W^*E \tag{11}$$

At this point an important difference arises between the analysis for the public and private sectors. The difference is that while private sector employees take both public and private wages as given and choose their consumption of public goods (and thereby private goods as well), public employees are assumed to have greater choice. We analyze three cases. First, analogous to the private employee optimum, we consider public sector optimization when the public sector wage (W_g) is fixed. Second, we assume that E_g is fixed, and solve to find the wage rate public employees would set if they had unlimited bargaining power. Third, we combine the first two to find the optimum when E_g and W_g are simultaneously determined. Unless otherwise indicated, in all cases it is assumed that the public sector employees acting in concert are aware of the effects of their policies on the tax price which they face (equation (11)).[12]

For a given value of W_g, the optimal level of public employment for public sector employees is found through a straightforward extension of the private employee optimization exercise by substituting (10) and (11) into (9) and maximizing with respect to E_g:

$$\frac{U_1}{U_2} = \frac{W_g^2}{W^*E} = P_g \tag{12}$$

or in the case of a CES utility function:

$$E_g = \left\{\left[\frac{b}{1-b}\right]^{1/(1+r)}\left[\frac{1}{W_g}\right]^{r/(1+r)} \cdot \left[\frac{W_g}{W^*E}\right]^{1/(1+r)} + \frac{W_g}{W^*E}\right\}^{-1} \tag{13}$$

[12]We are implicitly assuming that each public employee is naive in that he does not account for the fact that cuts in public output (i.e., E_g) will imply that there is a certain probability that a given public employee will lose his job and have his wage cut from W_g to W_p. Any attempt to account for this fully would substantially complicate the analysis. However, we might note that our discussion of bargaining strength in Section III does

where b is now the distribution parameter for public employees, but where the elasticity of substitution, $(1/1 + r)$, is the same as in the private case. Comparing (13) with (8), we see that the level of government employment desired by public employees is greater then that desired by private sector employees when

$$(14) \qquad \left(\frac{W_g}{W_p}\right)^r > \left(\frac{1-b}{b}\right)\left(\frac{a}{1-a}\right)$$

In the usual case where $r > 0$ (demands for public expenditures are price inelastic), this condition is met with $W_g = W_p$ and $a < b$ (the basic demand shift parameter is higher for public employees), when $a = b$ and $W_g > W_p$ (the income effect of higher public wages outweighs the substitution effect), or when some combination of the two conditions holds. Differentiating both (8) and (13) with respect to W_g, we see further that

$$(15a) \qquad \epsilon_p = \left(\frac{-1}{1+r}\right)\left(1 + r\frac{W_g E_g}{W^* E}\right)$$

$$(15b) \qquad \epsilon_g = \epsilon_p + \left(\frac{r}{1+r}\right)\left(1 - \frac{W_g E_g}{W^* E}\right)$$

where ϵ_g and ϵ_p are the price elasticities of demand for public and private employees, respectively. Again, when $r > 0$ the public employee demand elasticity will be less negative than that for private employees because the public wage W_g now has a dual effect, raising the price of public goods while also raising the income of public employees.

The second case we consider assumes that E_g is fixed and finds the wage that government employees would prefer if they had unlimited bargaining power. Obviously the fact that employees prefer this wage does not mean they will get it, but it is still fruitful to go through the case to see how the government wage rate and wage bill will tend to move as employees gain bargaining power.

The main limitation on the public sector wage rate in this case is the mobility of the private sector. We assume that there is at least one other accessible community and that public sector wage behavior is not identical in all jurisdictions.[13] To delineate the behavior of private employees further, we assume that each individual's level of utility attained depends not only on the level of public output and private consumption, but also on the characteristics of the community per se (for example, terrain, accessibility to relatives). To summarize this fact, we assume that each individual obtains a *quasi rent* from residence in the community with its unique characteristics. We also assume that there is a distribution of quasi rents, with the quasi rent for each individual given by A^i so that the actual level of utility of individual i is

$$(16) \qquad U_p^i = U_p(E_g, C_p; A^i)$$

where each quasi rent A^i is drawn from a known probability distribution and is strictly separable with respect to the other arguments of the utility function. Assume that each employee decides whether to move by comparing his level of utility U_p^i to the level of utility achievable in the best of all other jurisdictions, $U_p^{\prime\prime}$. If U_p^i is greater than or equal to $U_p^{\prime\prime}$, the private employee remains in the community. If U_p^i is less than $U_p^{\prime\prime}$, emigration occurs.

To simplify the analysis we assume that the parameters A^i are chosen so that private employees can be ranked in terms of the level of quasi rents that they achieve in the original jurisdiction. Specifically, we assume that if $A^i > A^j$ for employees i and j, then $U_p^i - U_p^{\prime\prime} > U_p^j - U_p^{\prime\prime}$. Define A^* to be the minimum value of A^i for all private employees residing in the original jurisdiction, as shown in Figure 1. All citizens with A^i greater than or equal to A^* reside in the community; all with $A^i < A^*$ do not.

Now consider the impact of a change in the public sector wage rate W_g. The gross income earned by public employees ($W_g E_g$) obviously rises, but since the model is open to trade as well as migration, the price of goods sold by the private sector is fixed, as are the gross receipts earned ($W_p E_p$). Hence if there is no

implicitly allow for control over the wage, and thus potential public sector benefits, to vary with the current size of the public sector.

[13]As long as it is possible to incorporate entirely new jurisdictions, as it still is in many states, this assumption will be technically fulfilled; and as long as there are many jurisdictions where relative public sector wages are low, the assumption will be fulfilled in practice.

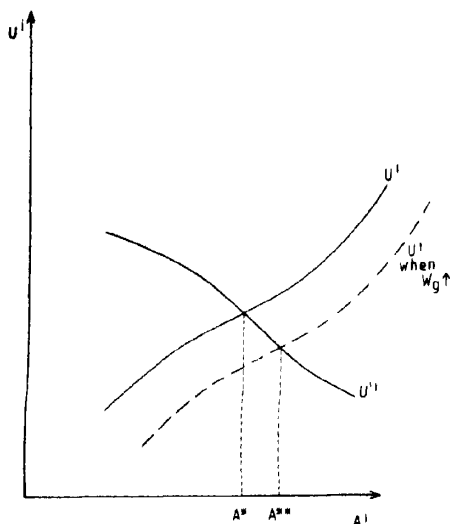



FIGURE 1. MOBILITY OF PRIVATE EMPLOYEES

migration W^*E will rise, but so will the tax rate on private income W_gE_g/W^*E. This increase in the tax rate implies that the consumption of private employees, given by (5) will decline, as will the utility of these employees. This outcome is shown graphically in Figure 1 as a downward shift in the U' schedule. The result is an increase in A^* to A^{**}, the emigration of all private employees with A' between A^* and A^{**}, a decline in E, and a decline in W^*E if the percentage decline in E exceeds the percentage rise in W^*.

In the more usual case the rise in W_g will lead to a decline in E_g (according to first-order conditions (6) or (12) above). Private employees then lose utility directly from the drop in E_g and indirectly if C_p falls (i.e., the tax rate rises). This in turn will happen whenever W_gE_g rises, or whenever demands for public services are price inelastic.[14]

[14]The tax rate $W_gE_g/(W_pE_p + W_gE_g)$ obviously rises initially whenever W_gE_g does. If W_gE_g falls, there is a mathematical possibility that this fall in the tax rate would lead to a large enough utility *gain* from the increased C_p to offset the utility loss from the fall in E_g and actually encourage in-migration. This case makes no logical sense, however, because if the private sector could actually benefit from a higher W_g, it would already be paying such a wage.

As a general matter, then, we can assert that for relevant values of W_g, E_g, and W^*, private employment and earnings (W_pE_p) will be negatively related to the public sector wage rate. If public sector employees are rational, they will take this mobility explicitly into account in determining their optimum wage levels, by recomputing the price of government services facing their members (equation (11)) as

$$P_g = \frac{W_g^2}{W^*E(W_g)} \tag{17}$$

where W^*E has been replaced by $W^*E(W_g)$, indicating that the tax base is a function of W_g.

It then becomes important to see how W^*E changes with W_g. There are two offsetting influences: a) the rise in W_g will alter the government wage bill W_gE_g directly; and b) the rise in W_g will reduce private employment E_p and lower the private wage bill and hence the tax base available to public employees. We can summarize these influences by defining the tax base elasticity with respect to public wages as

$$\eta = \frac{dW^*E}{dW_g}\frac{W_g}{W^*E} = \left(W_p\frac{dE_p}{dW_g} + E_g + W_g\frac{dE_g}{dW_g}\right)\frac{W_g}{W^*E} \tag{18}$$

where the first term in the parentheses is clearly negative, the second is clearly positive, and the final term is probably negative but could actually go either way depending on who is setting E_g and the value of r (see equation (15b)). The highest η could be is +1 (when $dE_p/dW_g = 0$ and when W_pE_p is very small relative to W_gE_g), but there appears to be no lower bound on the elasticity.

In the case we are considering, E_g is fixed, the third term in the parentheses drops out, and public employees maximize W_g given this level of E_g. This is tantamount to maximizing

$$C_g = W_g\left(1 - \frac{W_gE_g}{W^*E(W_g)}\right) \tag{19}$$

which (differentiating with respect to W_g) yields

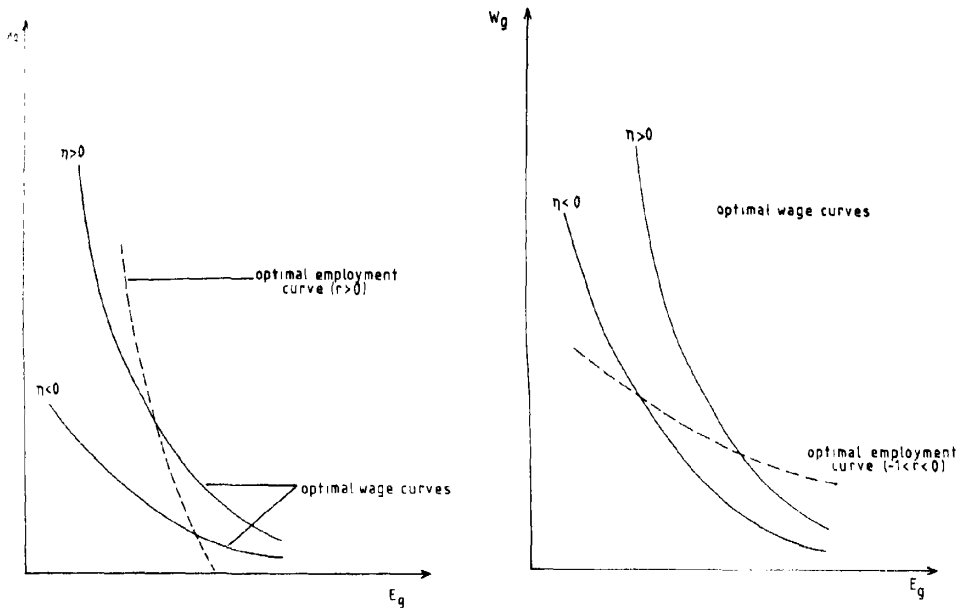



FIGURE 2. OPTIMAL WAGE AND EMPLOYMENT BEHAVIOR WITH MOBILITY

$$\frac{W_g E_g}{W^* E} = \frac{1}{2-\eta} \tag{20}$$

Hence the optimum solution is for employees to set the *entire* government budget as a share of the tax base. Other things equal, as the fixed level of government employment rises, the optimum wage declines. Further, when η is close to its highest attainable value of $+1$, the share of the government budget in the total tax base, $W_g E_g / W^* E = t$, is close to unity: public employees take almost all of total output. But as η becomes negative, the share of government declines gradually. Indeed, in the opposite extreme case where the set of A' for all private workers is sufficiently close to A^*, the private sector is so mobile and has so much bargaining power that the optimum public sector wage could get driven below the private wage. Again, we can rule out this outcome for logical reasons, because as soon as it begins to happen, public sector employees could also quit their jobs and work in the private sector. Hence (20) should be viewed as giving the upper bound on public wages and W_p as giving the lower bound.[15]

Next we consider the third case in which public sector employees are allowed to set E_g and W_g simultaneously. The first-order conditions (12) and (20) are the same, but now they must be solved simultaneously:

$$\frac{U_1}{U_2} = \frac{W_g}{E_g(2-\eta)} \tag{21}$$

This simultaneous equilibrium is summarized graphically in Figure 2, where in each panel the dotted line is the expenditure first-order condition (12) and the two solid lines are the

[15]With some algebraic manipulation, it can be shown that (21) is exactly equivalent to the first-order condition for the maximization of total tax revenue collected from the private sector, R_p. Since R_p can be written as tW_pE_p,

$$R_p = tW_pE_p = \frac{W_gE_g}{W^*E}(W^*E - W_gE_g) = W_gE_g\left(1 - \frac{W_gE_g}{W^*E}\right) = E_gC_g$$

When E_g is fixed, as in this case, the revenue taken from the private sector will be exactly proportional to the consumption of public employees, and maximizing C_g is the same as maximizing R_p.

possible wage first-order conditions (20), with the top line showing the situation where $\eta > 0$ and the bottom solid line showing the wage first-order condition when $\eta < 0$. As η becomes negative, the private sector is more mobile and the optimal wage for any E_g is shifted down. But the wage at which the ultimate intersection between the expenditure and wage condition takes place may be shifted either up or down, depending on public employees' price elasticity of demand for E_g (equation (15b)). If demand is inelastic (as in the left-hand panel), a community with a great deal of mobility (i.e., one in which η is negative) will have lower wages and a higher equilibrium level of government output than will a community with enough mobility so that η is exactly equal to zero. But if demand is elastic (as in the right-hand panel), a counterintuitive result arises: here the community with more mobility actually has a ***higher*** level of wages and a lower level of public employment than the community where not as much mobility is possible. Hence even though private sector mobility will constrain public employee wage demands at any given level of government employment, the full market equilibrium can lead to some surprising readjustments.[16]

III. Voting and Bargaining

In Sections I and II we described the determination by private and public employees of their desired levels of public employment. In this section we outline a resolution to the conflicting desires of private and public employees by introducing a simple majority-rule voting mechanism. In order to enrich the analysis we drop the assumption that the tastes of private and public employees are identical within groups, and instead assume that public and private employees have known distributions of tastes, which are in general different from each other. Thus, the analysis of desired levels of W_g and E_g now becomes an analysis of the levels desired by the median (or, more generally, the decisive) member of each group, and we assume that the chosen level of public output is equal to the desired level of the median voter.[17]

Consider first the case in which the public sector has a fixed amount of bargaining power and there is no mobility. Real wages in the public and private sectors are not necessarily equal—they could be higher in the public sector—but they are independent of current expenditure demands and the composition of employment in the community. In this case both sectors can be assumed to know W_g and W_p and vote for their optimal level of E_g, given for the private sector in equation (6) and for the public sector in (12) (where these now give the levels chosen by the decisive member of the respective groups). In general the two expressions imply different demands for public employment for three reasons: a) if relative wages differ, incomes differ and the income effect will lead to different public employment demands; b) if relative wages differ, tax prices differ and the substitution effect may lead to different demands; and c) the basic parameters of the utility function may not be the same.

If W_g is in excess of W_p, consideration a) would induce public employees to vote for larger public employment, consideration b)

[16]A plausible argument can be made that this case is unstable, while the inelastic demand case is stable. If, for example, moving expenditures were lowered so that η drops, W_g might begin falling as public employees recompute their first-order condition. This moves us toward equilibrium in the left-hand panel of Figure 2 and away in the right-hand panel.

[17]It should be noted that voting is only one of several political mechanisms (say, campaign contributions) which might affect the determination of the level of public employment, and some of the other mechanisms are likely to be more effective from a political point of view. In addition to political mechanisms, there is also the possibility of resolving conflicting demands by using the familiar economic condition for optimal provision of public goods. In the context of our model, this would require finding the public sector wage W_g such that with proportional income taxes the levels of E_g chosen by both public and private employees would be the same; i.e., the tax prices for each group would be the Lindahl prices. In the *CES* example, this condition can be examined by making inequality (14) an equality and solving for W_g as a function of the relevant parameters. In the case where public employees have some wage-setting power, the notion of a Lindahl solution seems to us to have little meaning, since the level of W_g is the instrument by which one party exploits another, and hence is hardly agreeable to the private employee.

FIGURE 3. DISTRIBUTION OF VOTERS FOR PUBLIC EMPLOYMENT

suggests smaller, and consideration c) is ambiguous, but probably dictates larger public employment. Regarding the latter, public employees might be sympathetic to the services they produce or have a job security motive where their own security depends on their seniority, which in turn would depend positively on the size of the public employee work force. As long as the substitution effect of higher public wages is small, and public employees have strong tastes for public output, they will vote for larger public employee work forces than will private employees, as shown in Figure 3.[18]

In light of these considerations we assume the median private sector voter will want a smaller public work force (E_g^p) than the median public employee (E_g^g). As long as E_g is less than half of E, there will more private sector workers (larger area) and the voting mechanism will result in an employment level between E_g^p and E_g^g, closer to the median of the private employee distribution, where the shaded area includes exactly one-half of the total voters and E_g^* represents the vote of the median voter of the entire resident population. Note that the median voter is either a private employee with a strong taste for public output (relative to that of other private workers) or a public employee with a weak taste for public output (relative to that of other public employees).

The question becomes somewhat more intricate when we realize that wage bargaining strength might also be endogenous. Once the public employee work force rises to some critical level, these employees become strong enough to influence elections, and it does not take an extreme cynic to imagine political candidates attempting to bribe unions into supporting the "right" party either with high wage increases or with the promise of institutional changes that facilitate future union bargaining.[19] We might assume that when the public sector is very small, public employee wages are set competitively at W_p, but that as E_g grows, public employees have progressively more power in moving to their desired wage given in (20). But there is a catch here: recall that (20) implies $W_g = W^*E/E_g(2 - \eta)$. As soon as E_g rises to the value of $E/(2 - \eta)$, the *optimal* wage desired by public employees is equal to the wage received in the private sector. Hence the maximum ratio of W_g to W_p will be achievable at a value of E_g between 0 and $E/(2 - \eta)$. This is shown in Figure 4 as a result of a political weighting of the competitive wage line ($W_g = W_p$) and the optimal wage line of expression (20). If $\eta > 0$, the range of E_g/E for which $W_g > W_p$ will extend past $E/2$, implying that public sector workers could simultaneously have a small enough sector to raise their wages and a large enough voting bloc to achieve their wage and expenditure objectives. But in general η is probably less than zero and the optimum economic level of E_g, that at which W_g is maximized, is almost certainly a good deal less than the optimum political level of E_g (where the voting influence is strong enough that workers can achieve their wage first-order condition).

The preceeding discussion is both similar to and different from Tullock's "Dynamic Hypothesis on Bureaucracy." Like Tullock, we have assumed that bureaucrats' political power is endogenous, and can be expected to

[18]Note that when utility functions are Cobb-Doublas ($r = 0$), the income and substitution effects exactly offset each other and the respective voting tendencies depend only on consideration *c*.

[19]That this actually happens in the world can be seen from a discussion of the history of union wage negotiations in New York City, see Raymond Horton and Gramlich.

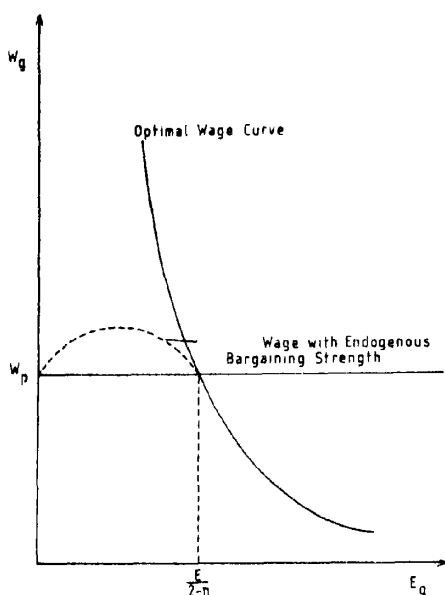



FIGURE 4. RELATIVE WAGES FOR THE PUBLIC SECTOR

be increasing in the fraction of the population which is employed in the public sector. But our analysis differs from his in one important respect: he suggests that bureaucrats will use an increasing fraction of their power to raise public sector *wages* as their employment rises. In our model this is not possible beyond some point, because the public employee's own *optimum* wage *decreases* as the level of public employment rises. Furthermore, the downward-sloping optimum wage curve is inherent in the technology of "exploitation" of the private sector—even if all of private sector income could be expropriated and there were no mobility (i.e., η approaches unity), the amount of such income available to each public employee must perforce be decreasing in the number of public employees. Thus, while we agree with Tullock that public employee *political power* will increase in the level of public employment, the ability of public employees to convert that power into high wages will be attenuated both by the mobility of the private sector and by the simple arithmetic of "dividing up the pie."

Putting things in this light implies that even a strong public sector union should ultimately be controllable. Since the total government budget desired by public employees is constrained by private sector mobility, public sector monopolies interested in maximizing the individual employee's wage will find it economically optimal to aim at a share of total employment that should be substantially less than $E/2$. As long as voting participation rates are the same among public and private sector employees, this assures that the public sector worker-voters will remain a minority. Moreover, there are two other constraints on behavior which temper public employee power even further: a) public employees themselves will be schizophrenic in the sense that their own public-private spending desires (equation (12)) may conflict with the E_g solution of Figure 4; and b) private sector employees may learn that increases in E_g increase the risk that the public sector can move toward its optimal wage curve, and these private employees may take this as another constraint in their own voting.

IV. Implications

At this point it may be helpful to return to the original message and pull together what this model has to say about it. The original fear was that as public servants are hired, they would become a steadily more dominant electoral force, elect politicians who would grant them steadily more market power, and the upshot of the rising levels of E_g and W_g would be steadily greater public budgets and higher tax rates. The public sector becomes dominant and the private sector is struggling, maybe unsuccessfully, for survival.

This model tries to put these fears in perspective by treating them analytically. The major conclusion is that as long as the private sector retains its right to leave the community, and as long as all communities are not suffering high and rising tax rates, public sector monopolies ought to be kept in check by the simple optimizing behavior of these same public employees.[20] It shows further

[20]This does not mean to imply that public sector unions always perform their optimization calculations correctly. A plausible interpretation of the recent history of New York City, for example, would be that the municipal unions' estimate of η was well above the true value of that

that the optimum public wage depends inversely on the size of government employment. Adding public employees may increase the political strength of unions but it also decreases the wage they will desire to attain. Finally, on the demand side, both public and private voters in their expenditure demands will also be price sensitive. As public employees try to enforce higher levels of public sector wages, private voters will force down the share of public employment, which should reduce even more the chance the public sector monopolies will be able to raise wages and exploit the private sector.

REFERENCES

J. L. Barr and O. A. Davis, "An Elementary Political and Economic Theory of Local Governments," *Southern Econ. J.*, Oct. 1966, *33*, 149–65.

T. C. Bergstrom and R. P. Goodman, "Private Demands for Public Goods," *Amer. Econ. Rev.*, June 1973, *63*, 280–96.

R. M. Bird, "Wagner's 'Law' of Expanding State Activity," *Publ. Finance*, No. 1, 1971, *26*, 1–26.

T. E. Borcherding and R. T. Deacon, "The Demand for the Services of Non-Federal Governments," *Amer. Econ. Rev.*, Dec. 1972, *62*, 891–901.

———, **W. C. Bush, and R. M. Spann,** "The Effects of Public Spending of the Divisibility of Public Outputs in Consumption, Bureaucratic Power, and the Size of the Tax-Sharing Group," in Thomas E. Borcherding, ed., *Budgets and Bureaucrats: The Sources of Government Growth*, Durham 1977.

J. Buchanan, "Why Does Government Grow?," in Thomas E. Borcherding, ed., *Budgets and Bureaucrats: The Sources of Government Growth*, Durham 1977.

——— **and G. Tullock,** "The Expanding Public Sector: Wagner Squared," *Publ. Choice*, Fall 1977, *23*, 147–50.

---

parameter. Moreover, in constantly comparing their own wage levels to wage levels in other cities where the size of the public employee work force is smaller, these employees are hurting both the private sector (because the price of public services is raised) and ultimately themselves (because the tax base is falling).

W. C. Bush and A. T. Denzau, "The Voting Behavior of Bureaucrats and Public Sector Growth," in Thomas E. Borcherding, ed., *Budgets and Bureaucrats: The Sources of Government Growth*, Durham 1977.

R. Craswell, "Self-Generating Growth in Public Programs," *Publ. Choice*, Spring 1975, *21*, 91–98.

R. G. Ehrenberg and G. S. Goldstein, "A Model of Public Sector Wage Determination," *J. Urban Econ.*, July 1975, *2*, 223–45.

E. M. Gramlich, "The New York City Fiscal Crisis: What Happened and What is to be Done?," *Amer. Econ. Rev. Proc.*, May 1976, *66*, 415–29.

Raymond D. Horton, *Municipal Labor Relations in New York City: Lessons of the Lindsay-Wagner Years*, New York 1973.

Norman MacRae, *America's Third Century*, New York 1976.

William A. Niskanen, "Bureaucrats and Politicians," *J. Law Econ.*, Dec. 1975, *18*, 617–43.

———, *Bureaucracy and Representative Government*, Chicago 1971.

Joseph A. Pechman and Benjamin A. Okner, *Who Bears the Tax Burden?*, Washington 1975.

J. H. Pencavel, "The Demand for Union Services: An Exercise," *Ind. Labor Rel. Rev.*, Jan. 1971, *24*, 180–90.

M. W. Reder, "The Theory of Employment and Wages in the Public Sector," in Daniel S. Hamermesh, ed., *Labor in the Public and Nonprofit Sectors*, Princeton 1975.

S. P. Smith, "Government Wage Differentials," *J. Urban Econ.*, July 1977, *4*, 248–71.

G. R. Tullock, "Dynamic Hypothesis on Bureaucracy," *Publ. Choice*, Fall 1974, *19*, 127–31.

R. B. Victor, "The Effects of Municipal Unionism on Wages and Employment," unpublished doctoral dissertation, Univ. Mich., 1977.

R. E. Wagner and W. E. Weber, "Wagner's Law, Fiscal Institutes, and the Growth of Government," *Nat. Tax J.*, Mar. 1977, *30*, 59–68.

J. Wanniski, "Taxes, Revenues, and the 'Laffer Curve'," *Publ. Interest*, Winter 1978, *50*, 3–16.

The Monetary Approach to Official Reserves and the Foreign Exchange Rate in France, 1962–74: Some Structural Estimates

By JACQUES MELITZ AND HENRI STERDYNIAK*

Any test of the monetary approach centered on the period of fixed exchange rates would now be predominantly of historical interest. At the time of this study, however, experience with flexible exchange rates was still too short to permit concentrating econometric analysis exclusively on this more recent system. Caught in this net, we have attempted an analysis of official reserves and the foreign exchange rate in France covering both fixed and flexible exchange rates, that is, thirty-nine quarters of fixed rates, 1962.1–1971.3, and thirteen quarters of flexible rates, 1971.4–1974.4. The cost is the presence of errors in our simultaneous equation estimates of the exchange rate during the period of fixed rates. But the benefit is an econometric analysis founded on fifty-two observations, and yet covering three years of flexible rates. Since the errors in the estimates of the exchange rate under fixed rates are quite moderate, the cost would seem to be worth the benefit.

The most important characteristic of our work is the use of a detailed structural model of bank credit and money in testing the monetary approach. The early tests of this approach simplified the structure of the monetary system to the utmost and considered the domestic source component of the reserve base (or the total base minus official foreign reserves) and the money multiplier as exogenous.[1] But there is really no logical basis for these restrictive assumptions. The monetary approach states that the demand and supply of money in a "small" country together determine 1) money, and 2) official reserves or the foreign exchange rate or the attainable combinations of the two, depending upon fixed, floating, or managed exchange rates. Nothing but a correct specification of the conditions for monetary equilibrium can provide a basis for testing this proposition.

We also deviate from the tendency in the literature on the monetary approach to suppose that any convenient measure of money will do. Based on this attitude, there have been many tests of the monetary approach using simply the reserve base as the measure of money, even though this aggregate, consisting of currency plus an arbitrary fraction of deposits, is inappropriate in analyzing the monetary behavior of firms and households.[2] In justifying this measure in a well-known econometric work, Pentti Kouri and Michael Porter merely say: "The essential features of the model [would not be] substantially changed by incorporating a more complete banking system" (p. 448). But not only does this fail to meet the criticism, it also neglects the fact that the monetary approach can give rise to conflicting estimates of changes in official reserves and the exchange rate depending on the money measure.[3] There is no way of assessing the seriousness of this last objection without testing. In this work we shall examine the extent to which varying and tenable money measures in France yield convergent results.

In spite of these deviations from the literature, we may be said to adhere to a strict

*Institut National de la Statistique et des Études Économiques, Paris.

[1]See, for example, the works of J. Richard Zecher, Hans Genberg, and Donna Bean.

[2]Harry Johnson showed the way in his seminal article

[3]Nonetheless, Kouri and Porter invite us to take seriously their point estimates of the extent to which changes in official reserves neutralize domestic monetary policy actions in four different countries under fixed exchange rates. With respect to the issue of choice of money aggregate, see Rudiger Dornbusch, p. 257, fn. 1. One important article anticipating Kouri and Porter's work about one decade earlier is by Jürg Niehans.

version of the monetary approach. Only monetary influences enter into our analysis. We take real income as given and ignore the balance of trade and the terms of trade. Any use of the trade balance as a datum in determining the flow of official reserves would limit the analysis to the explanation of movements in reserves corresponding to foreign capital movements. By eschewing information about absorption or the terms of trade—anything to do with imports or exports of *commodities* as such—we avoid this restriction and examine a particularly strong version of the monetary approach, in accordance with Harry Johnson's influential paper.

Following Kouri and Porter, our effort might also be expected to show the extent to which movements in official reserves offset the impact of policy changes in the reserve base. However, there are no open market operations in France; the basic monetary instrument is bank rate, or more precisely, the daily price of refinancing at the Bank of France (*le taux de l'argent au jour le jour*). Thus the only question we can investigate which corresponds to Kouri and Porter's problem is the extent to which a fiscal deficit is offset by an outflow of official reserves. In fact, we find this offset to be very high, on the order of 70 percent within a year (a reasonable figure since practically all fiscal deficits were financed directly through the central bank during most of the study period). The offset also has only diminished slightly—by about 4 percent—since the abandonment of fixed rates in late 1971. In general, the regime of flexible rates has brought little change in the magnitude of the influences on money and official reserves in France. There is simply a new and high sensitivity to the dollar-deutschmark exchange rate. We shall conclude, on the basis of the analysis, that the monetary approach is quite useful, but does not explain enough to warrant exclusive reliance upon it.

I. The Model

The general underlying conception of money-stock determination in this work has important links to Karl Brunner and Allan Meltzer (1966, 1968). On the one side, commercial banks demand earning assets; on the other, firms and households supply such assets to the banks. There results a quantity of earning assets and a price. The quantity of commercial bank earning assets, together with the banks' desired excess legal reserves, their desired borrowing from the central bank, the legal reserve requirement, and an accounting identity yield commercial bank deposits. An equilibrium relationship between bank deposits on the one side and government monetary debt on the other, then permits deriving the money stock if money is viewed as the sum of commercial bank deposits and the monetary liabilities of the government (or as the sum of these deposits plus currency, postal checking deposits, and state savings deposits). If money should be understood differently, then some other deposit ratios would enter the analysis.

Adapting this fundamental framework to France is easy in one respect, complicated in another. What makes it simple is that the earning assets of French commercial banks consist entirely of loans, and not securities. The banks demand no securities since they can use eligible paper for discount at the central bank (*effets mobilisables*) to satisfy all of their demand for interest-yielding liquidity.[4] Consequently, there is no need first to determine the supply and demand for bank loans, then separately to determine the demand for securities by the banks and the supply of securities to them, and finally to aggregate in order to obtain the demand and supply of bank earning assets. The complication in France is that the demand for commercial bank loans cannot be distinguished from the demand for loans from other financial institutions in the sphere of the Treasury.[5]

[4]Prior to 1967, commercial banks were required to hold some government securities. We have taken this into account in the econometric analysis, though we shall ignore it in the discussion.

[5]The most important public credit institutions under Treasury control include the Bank of Deposits and Consignations, which administers all of the funds collected through the savings banks, the Fund for Economic and Social Development, and the Land Bank or Crédit Foncier. It should be carefully noted, however, that there are some publicly owned banks in France that

Our precise formulation of money-stock determination in France is easily understood in view of these considerations. Our model is shown in the following equations. (The symbols are defined in Appendix A.)

Commercial bank reserves

$$(1)\quad R = R_r + R_e$$

Bank credit

$$(2)\quad C = C_b + C_t + RD$$

Commercial bank net assets

$$(3)\quad A = C_b + R_e$$

Commercial bank accounting identity

$$(4)\quad C_b = D_b - R$$

Reserve base

$$(5)\quad B = D_{po} + D_{rb} + N + R$$

Base money

$$(6)\quad B_m = B - R$$

Money in the sense M_4

$$(7)\quad M_4 = D_b + B_m$$

Demand for bank credit

$$(8)\quad \frac{C}{P} = \frac{C}{P}(\overset{-}{i}, \overset{+}{i^*}, \overset{+}{dP_a}, \overset{+}{Y_p}, Y_{tr})$$

Supply of bank credit

$$(9)\quad C = C(\overset{+}{i}, i_1, \lambda, r, \overset{+}{F}, \overset{+}{CFD}, \overset{\pm}{Y_p}) + C_t$$

Commercial bank demand for net assets

$$(10)\quad A = A(\overset{+}{i}, i_1, \lambda, r, \overset{+}{F}, \overset{+}{CFD}, \overset{\pm}{Y_p})$$

Equation of the ratio of money to net assets

$$(11)\quad \frac{M_4}{A} = \frac{1}{1-r} + \psi_4(i, \overset{+}{i_1}, \overset{+}{\lambda}, \overset{+}{r}, \overset{\pm}{Y_p})\frac{1}{1-r}$$

where $\psi_4 = B_m/D_b$

Demand for money

$$(12)\quad \frac{M_4}{P} = \frac{M_4}{P}(\overset{+}{Y_p}, dP_a)$$

Reaction function of the monetary authorities

$$(13)\quad \frac{FF}{\$} = \frac{FF}{\$}(\overset{-}{FX}, \overset{+}{\frac{DM}{\$}}X, X) + TAR$$

Price level equation

$$(14)\quad P = P(\overset{+}{\frac{FF}{\$}}) + P_x$$

Exogenous variables: C_t, CFD, $DM/\$$, i^*, i_1, P_x, dP_a, r, TAR, X, Y_p, Y_{tr}, λ.

Endogenous variables: A, C, F, $FF/\$$, i, M_4, P.

As can be seen from this system of equations, we introduce a demand and supply of bank loans inclusive of the loans of the financial institutions in the Treasury sphere as well as those of the commercial banks (see equation (2)). These represent the aggregate demand and supply of bank credit in the analysis (equations (8) and (9)). Then we use a separate behavioral relation (equation (10)) to determine the commercial banks' contribution to this supply of bank credit, minus their rediscounts and plus their demand for excess legal reserves, or the banks' "demand for net assets." A fourth equation, (11), reflecting the joint behavior of commercial banks, households, and firms, sets the ratio of money to commercial bank net assets, or equivalently, given the legal reserve requirement, the ratio of money to commercial bank deposits. Money is a very broad concept here, inclusive of all privately held government debts with a fixed nominal price (previously referred to as government monetary liabilities), that is, postal checking accounts and deposits at the savings banks as well as currency plus all commercial bank deposits. The advantage of this broad measure of money, M_4, is that the function ψ_4 ( ) of equation (11) consequently refers to the competitive position of the commercial banks relative to the government in the issue of deposits and currency. The previous four behavioral equations together determine bank credit, the interest rate on bank credit, commercial bank net assets, and money in the sense M_4.

In line with the monetary approach, the additional presence of a demand for money (equation (12)) serves to determine official

should be viewed as ordinary members of the commercial banking system. As all students of the French monetary system agree, these are the three nationalized commercial banks (Crédit Lyonnais, Banque Nationale de Paris, and Société Générale), the branch system of Agricultural Banks, and the branch system of Popular Banks. For a few other prominent studies of the French system, see the works of Antoine Coutière and Jacques-Henri David reprinted in the Commissariat Général au Plan. Also see André Fourçans, and Patrick Artus.

reserves under fixed rates or the foreign exchange rate under floating rates. Equation (13), the official reaction function, permits us to view official reserves and the foreign exchange rate as simultaneously determined. The final equation, (14), ties the price level to the foreign exchange rate, and thereby admits the influence of the foreign exchange rate on the demand for money and the demand for credit. Without this last influence, the impact of the foreign exchange rate in the model would depend solely on the preoccupation of the authorities with the value of the franc in setting their net foreign reserves.

The signs above the variables of the fundamental equations (8)–(14) relate to the partial derivatives. A few points in the formulation of these equations may be clarified. The cost of foreign borrowing in the demand for credit, equation (8), is simply the foreign interest rate i^*, and not the sum of this rate and the forward premium on the franc, because earlier tests indicated that this forward premium is statistically insignificant. In equation (9), the cumulative fiscal deficit (*CFD*) is an argument, rather than the Treasury source component of the base, because the Treasury source component implicitly depends on the amount of base money, $M_4 - D_b$ (or $B - R$), and cannot be viewed as exogenous. For instance, the Bank of Deposits and Consignations in the Treasury sphere holds a large volume of securities which, based on the model and all evidence, varies with the demand for base money. Permanent real income Y_p is also an argument in equation (9) because it may affect ψ_4 in equation (11). If Y_p should lower this ratio, for example, thus raising the commercial banks' share of the market for issues of deposits, it would obviously raise the supply of credit. Equation (9) thus is a partially reduced form. This is also true of equation (10), which equals $C(\ )$ in the supply of credit minus the desired value of net borrowed reserves $RD - R_e$.[6]

The signs of i, i_1, and r in the functional relation $\psi_4(\ )$ of equation (11) rest on the hypothesis that any rise in the profitability of asset expansion will induce the commercial banks to offer more services to their depositors. A rise in λ, on the other hand, will raise ψ_4, even though necessitating more commercial bank services (λ being a certain legal disadvantage of the commercial banks relative to the savings banks with respect to the amount of interest they can pay and certain tax regulations). Permanent income Y_p possibly affects ψ_4 because the income elasticity of the private demand for commercial bank deposits may differ from the income elasticity of the private demand for base money.

There are two fundamental omissions in equation (12), the demand for money: the foreign interest rate i^* (with or without the forward premium on the franc); and the money rate on domestic bonds. Both omissions rest on econometric considerations. We were unable to find a significant negative coefficient of i^* in the simultaneous equation estimates of the demand for money (though we succeeded in doing so in the single equation estimates), and we also found that anticipated inflation dominated the money rate on domestic bonds in the estimates of the demand for money (see Melitz, 1976a).

Equation (13) says that the foreign exchange rate $FF/\$$ is equal to the target rate under fixed exchange rates, prior to 1971.4, and since then equal to

$$\frac{FF}{\$} = \frac{FF}{\$}\left(F, \frac{DM}{\$}, 1\right)$$

The hypothetical negative sign of F in the function means that in case of a downward pressure on the franc, the French authorities prefer a combination of a rise in $FF/\$$ and a fall in F to merely one or the other. The $DM/\$$ in the reaction function refers to the influence of the Common Market "snake." The separate presence of F implies the possible French absence from the snake. If the fit of the equation is good, the official presence or absence of France in the snake can be explained by the equation.

As regards the price level equation, (14), the exogenous component P_x refers partly to the influence of foreign prices of traded goods, partly to the influence of French prices of nontraded goods (since these prices may be

[6] For a detailed discussion of the relation between $C(\ )$ and $A(\ )$, see Melitz (1976b, Part I, pp. 56–58), which contains an econometric study of the same model of the French monetary system with F and $FF/\$$ as exogenous.

independent of the foreign exchange rate to some extent), and partly to changes in the relative weights of the two.

The only uncomfortable aspect of the system is the view that anticipated inflation dP_a is exogenous. But a very delicate problem confronted us in this respect. To have treated dP_a differently, or as dependent on P, thus $FF/\$$, would have been to allow the errors in the fitted values of $FF/\$$ to distort the fitted values for anticipated inflation as well as those of P during the period of fixed exchange rates, covering three-quarters of our observations. The distortion then would have persisted during the early part of the period of flexible rates because of distributed lag effects of P on dP_a. It seemed to us wiser, therefore, to accept the element of simultaneous equation bias (affecting only the tail end of the period) which is inherent in the treatment of anticipated inflation as exogenous.

We studied three other measures of money besides M_4, namely, M_1, M_2, and M_3; but space limitations will confine the discussion of the econometric results to one, M_1. Therefore we shall specify the adaptation of the model only to this one other money measure (though the adaptation to the other two, M_2 and M_3, will become apparent).[7]

Let M_1 be privately held currency plus demand deposits, or

$$M_1 = M_4 - D_{sb} - D_{bs} \tag{7'}$$

where D_{sb} are the savings deposits with the savings banks, D_{bs} those with the commercial banks, and $D_{sb} + D_{bs}$ total savings deposits. Instead of equation (11), we then have

$$\frac{M_1}{A} = \frac{1}{1-r} + \psi_1(\quad)\frac{1}{1-r} \tag{11'}$$

where $$\psi_1 = \frac{D_{po} + N - D_{bs}}{D_b}$$

The problem then is to show, in line with $\psi_4(\quad)$:

$$\psi_1 = \psi_1(\overset{-}{i}, \overset{+}{i_1}, \overset{+}{\lambda}, \overset{+}{r}, \overset{\pm}{Y_p})$$

A rise in i will lead commercial banks to try to attract deposits on previous grounds, thus lowering the ratio of currency plus postal checking accounts to their deposits. Since law permits these banks to increase the interest they pay on certain types of savings deposits, but not to pay any interest on checking accounts, their efforts to raise their deposits will partly provoke a rise in the ratio of their savings deposits, D_{bs}/D_b, which will then reinforce the fall in ψ_1. The positive signs of i_1 and r in $\psi_1(\quad)$ follow on the same ground. As regards the sign of λ, deposits at the savings banks are closer substitutes for savings deposits at the commercial banks than they are for either currency or checking accounts. Consequently a rise in λ will raise $(D_{po} + N)/D_b$, lower D_{bs}/D_b, and therefore raise ψ_1.

The demand for money in the sense M_1 may be stated as

$$\frac{M_1}{P} = \frac{M_1}{P}(\overset{+}{Y_p}, \overset{-}{dP_a}, \overset{-}{i_d}) \tag{12'}$$

where i_d is the interest rate on savings deposits. Equations (11′) and (12′) then can substitute for (11) and (12) in the case of M_1.

II. The Method of Estimation

There are three basic problems in estimating the system of equations (8)–(14). The first concerns the missing series P_x in the price level equation. We handled this difficulty by identifying P_x with the sum of the constant and the disturbance term $(a_o + \mu)$ in the equation

$$P = a_o + a_1\frac{FF}{\$} + \mu \tag{14'}$$

In other words, we broke up P between P_x and $a_1FF/\$$ on the basis of a one-stage least squares estimate of equation (14′), and next substituted $P_x + a_1FF/\$$ for P in equations (8) and (12). In this way, we reduced the system to six equations before engaging in simultaneous equation estimation. This poses an obvious risk of simultaneous equation bias

[7] M_2 is currency plus total commercial bank deposits, and M_3 is currency plus demand deposits plus savings deposits with either the commercial banks or the savings banks that are low-interest yielding and without any maturity date.

in the estimate of a_1. But the procedure has the fundamental merit of assuring that the differences between the observed and estimated values of P in the simultaneous equation estimates derive strictly from differences between the observed and estimated values of $FF/\$$. There would be no point in admitting errors in the estimates of P stemming from any other source in a test of the monetary approach as such. Our single equation estimate of equation (14′) involves a highly significant coefficient a_1 (with a t-statistic of 4.7).

Our second estimation problem is that several of our equations are not linear with respect to all of the endogenous variables. This is true for equation (8) because of the expression $C/(P_x + a_1FF/\$)$, for equation (12) because of the expression $M/(P_x + a_1FF/\$)$, and for equation (11) because of the expressions M/A and $i/(1 - r)$. The problem could not be resolved by stating all of the equations in *log*-linear form, since the effects of F and CFD in the supply of credit and the demand for net assets (equations (9) and (10)) are genuinely linear rather than *log*-linear. Moreover, the linear form of the demand for credit in France happens to yield substantially better fits than the *log*-linear form. We therefore formulated all of the equations linearly, using linear approximations to the four previous expressions.

Our third problem involves the difficulty of treating official reserves in the reaction function (equation (13)) as endogenous for only part of the observation period. Using FX instead of F in the reaction function does not resolve this problem fully, as it may seem to, since we cannot consider FX (which is zero until 1971.4 and then becomes equal to official reserves) as an endogenous variable as distinct from F. Hence, we were compelled to substitute $F - F(1 - X)$ for FX in equation (13), interpret F as endogenous, $-F(1 - X)$ as exogenous, and constrain both F and $-F(1 - X)$ to have the same coefficient. This procedure works perfectly with flexible exchange rates when $X = 1$ and therefore $-F(1 - X)$ is zero. But under fixed rates when $X = 0$, the sum of F and $-F(1 - X)$ does not yield zero as we would like, since F is the estimate of official reserves, $-F(1 - X)$ the negative of the observed value of official reserves, and $F - F(1 - X)$ therefore the error in the estimate. This error then becomes a factor distorting our estimates of the exchange rate in the period of fixed rates. We found this distortion impossible to avoid; in other words, we found no way to suppress the simultaneity of the estimates of F and $FF/\$$ for any subsection of the study period. Fortunately, however, the errors in the estimates of $FF/\$$ are only moderate under fixed rates. Two-thirds of them are within 2 percent and with only one close exception all of the rest are within 5 percent.

As regards the structure of the distributed lags, we admitted a fairly short adaptation period in all of the equations except for the demand for credit. Earlier experiments with Almon lags showed that there was no point in deviating from a simple Fisherine assumption of a linear decay in influence. Accordingly, we imposed a linearly declining pattern of distributed lags everywhere lasting four quarters, except in the demand for credit where they last twelve, and of course, in the reaction function where the adjustment supposedly takes place in a single quarter. There are a few exceptions: Y_{tr} in the demand for credit and r in equations (25) and (26) are all supposed to exert their full effect in a single quarter.

We also introduced a dummy variable in all six equations for the third quarter of 1969, when France devalued by 10 percent under fixed rates. The relevant theory of the determination of F does not apply in that quarter. Further, it proved useful to inject the dummy variable X in all of the equations, not only (13). The variable X always has a highly significant coefficient and its presence diminishes the standard errors of some of the other coefficients without much affecting their levels. The simplest interpretation is that X compensates for a bias issuing from the errors in our estimates of $FF/\$$ under fixed exchange rates.

The computer program at our disposal (Clifford Wymer's well-known *RESIMUL*) enabled us to estimate the system with full-information maximum likelihood (*FIML*)

subject to all of the desired linear restrictions.

III. The Econometric Analysis

The *FIML* estimates are shown in Appendix B, together with explanatory material relating to measures, data, statistical magnitudes, and the meaning of test statistics. We simplify to the utmost by ignoring the dummy variable for 1963.3 (which proved utterly insignificant), the dummy variable for flexible exchange rates in the first five equations of the Appendix, and the insignificant coefficient of r (more precisely $r/(1-r)$) in the ψ function (which we set at zero in the estimates). The Appendix shows the predetermined coefficient of the influence of the exchange rate on the price level to be 0.06 (with a t-statistic of 4.7 in the single equation estimate, as we mentioned before). This corresponds to an exchange rate elasticity of the price level of 0.2 at the means, which is not as small as it may seem, since the effect is over a single quarter.[8]

The coefficient of the influence of Y_p on M_4/A (in equation (11)) is negative and significant, which theoretically implies a positive effect of Y_p in the supply of bank credit and the demand for net assets. However, we encountered a considerable problem of multicollinearity between Y_p, F, and CFD in these two equations. With Y_p as a factor in both equations, the coefficients of F vary greatly from one measure of money to the next, they are generally unreliable, and the reduced form is largely shorn of interest. Thus, we report only on the *FIML* estimates without Y_p in equations (9) and (10).

These estimates, or the ones in Appendix B, are quite satisfactory on the whole. The only major exception is the coefficient of the interest rate i in the demand for bank credit in the model with M_1, which is insignificant and of the wrong sign. The coefficient is negative and significant, however, in all of the estimates with the other money measures (M_2 and M_3 as well as M_4). Another bothersome result is the insignificance of the legal reserve requirement r (which first appeared in France only in 1971) in the supply of credit. But this is less embarrassing than it may seem, since a rise in r necessitates more refinancing (RD) at any given level of deposits and in this respect tends toward a positive sign of r in this equation. On the other hand, the negative effect of r in the demand for net assets, about which there is no theoretical ambiguity, is amply confirmed in the estimates.

The enormous t-value of $DM/\$$ in the reaction function is not surprising because of the occasional official French presence in the European Economic Community (EEC) snake. The real mark of the success of the equation, therefore, is the equally high t-value of the coefficient of F. Since the coefficient of $DM/\$$ does not differ significantly from one, the conclusion is that, *ceteris paribus*, the French authorities try to keep the DM/FF exchange rate about the same. However, being that F clearly is another official consideration, the periods when France is in the EEC snake evidently correspond to those when keeping the franc moving in perfect step with the DM relative to the dollar requires little change in reserves.

We do not provide the R^2 for $FF/\$$ in Appendix B since, as explained there, this statistic would have no interest: three-quarters of the observations in the study period concern fixed exchange rates. Instead, we have constructed a special R^2 relating to $FF/\$$ and F during the period of flexible exchange rates (see Appendix B). These R^2s, in combination with the others in the appendix, yield two basic conclusions: first, that the model explains F and $FF/\$$ less well than the other four endogenous variables;[9] second, that

[8]With only thirteen quarterly observations at our disposal, we tried but could not find any statistically significant pattern of distributed lags in the influence of $FF/\$$ on P. Our 0.2 estimate of the exchange rate elasticity is in line with the estimates of Dornbusch and Paul Krugman, pp. 568–73, for various countries, including France.

[9]There may be some hesitation about this conclusion on the ground that all of our estimates are interdependent; therefore the estimates of F and $FF/\$$ could even be the source of the high R^2s for the other variables. But we know from *FIML* estimates of equations (8)–(11) alone for the same study period (see Melitz, 1976b, 1977) that the quality of our estimates of C, i, A, and M_4 or M_1 owes nothing to the fitted values of P, F, and $FF/\$$. This would then support the statement in the text.

F (billion francs)

···· Predicted in case M_4
---- Predicted in case M_1
—— F observed

40
30
20
10

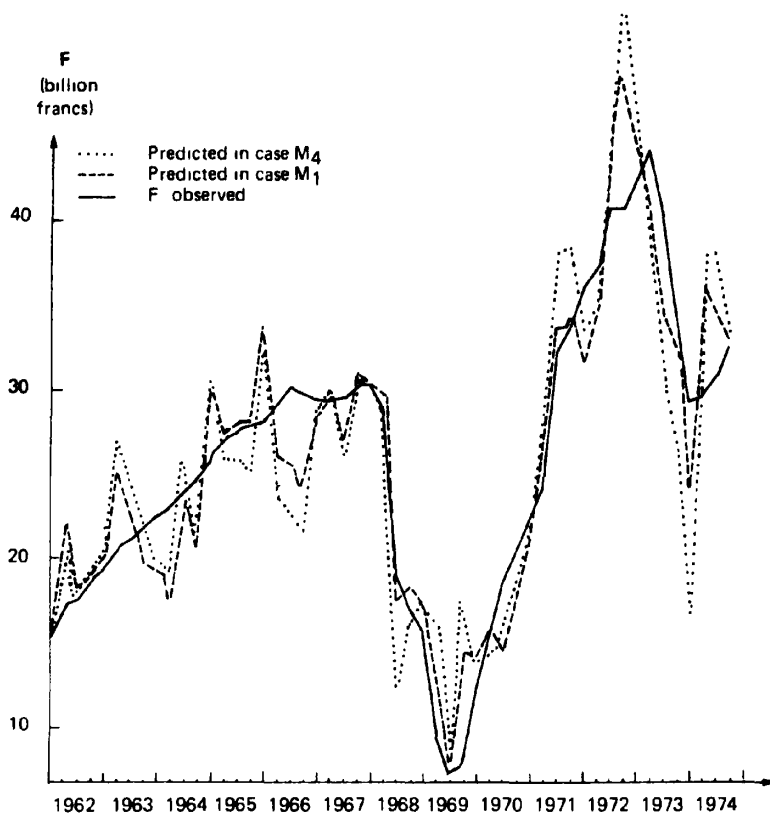
1962 1963 1964 1965 1966 1967 1968 1969 1970 1971 1972 1973 1974

Coefficient of correlation in 1971,4-1974,4 (r)=0,57 in the case of M_4 and 0,74 in the case of M_1

FIGURE 1

the model explains these two variables better in the case of M_1 than M_4.

It may be seen from Figures 1 and 2 that the profiles of the errors in the estimates of F and $FF/\$$ are basically the same for M_1 and M_4. The model does not explain F particularly well in either case during the calm of 1962.1–1968.1 when F grew steadily under fixed exchange rates (see Figure 1). The fitted values define some fictitious, sharp, tooth-edged movements around the basic upward trend during this time. The model is very successful, however, in tracking the sharp downswing in F starting with the political disturbances of May–June 1968 and the subsequent sharp upswing in F following the devaluation of the franc in 1969.3. It also anticipates the downswing of F of 1973.3 by one quarter and captures the timing of the upswing after the oil crisis in mid-1974. With respect to $FF/\$$, the model traces the appreciation of the franc from 1971.4 to 1973.2 with a certain lack of precision (see Figure 2). But it captures the timing of the depreciation of 1973.3 and the timing of the subsequent appreciation of 1974.1. The performance of the model therefore is reasonable at the turning points and during the turbulent episodes.

Table 1 provides the reduced-form coefficients of the influences on F, $FF/\$$, and M. These coefficients, like the structural ones of

FIGURE 2

Appendix B, refer to cumulative impacts over time. All of the effects are distributed over three years, but attain 85 percent of their cumulative value within two years in some cases—namely, those involving the variables entering in the demand for credit (P_x, i^*, dP_a, and Y_p)—and more than 95 percent of their value within a single year in the other cases ($DM/\$$, i_1, λ, CFD, C_t). According to the model, variables which do not appear in the demand for money cannot affect the quantity of money under fixed rates. This explains all the null effects on M_1 and M_4 when X is zero in Table 1.

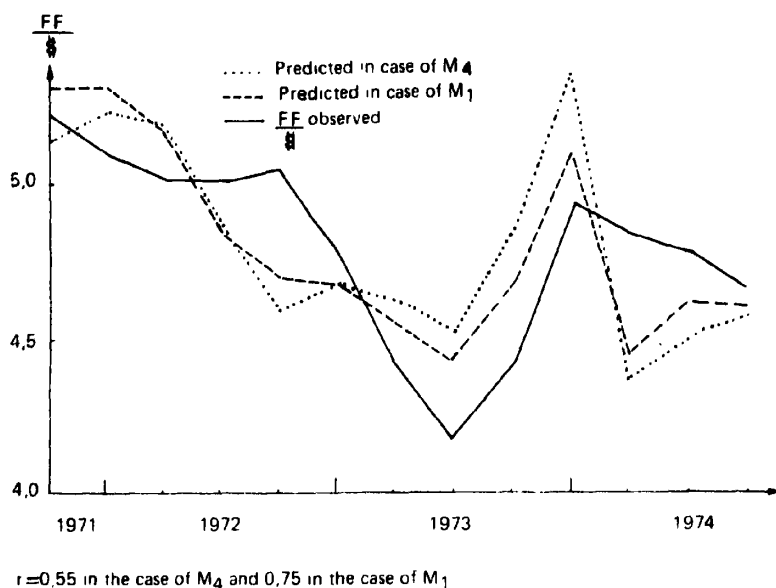



Table 1 implies that the rate of intervention of the Bank of France, i_1, has perverse effects on money, reserves, and the franc. This can only mean, contrary to theory, that the positive effect of i_1 on M/A dominates the negative one on A. An important consideration in this regard is the large competitive presence of the French government in the issue of deposits via the system of postal checking accounts and the savings banks. As the commercial banks reduce their services to depositors subsequent to a rise in i_1, there is therefore wide scope in France for a rise in deposits with the government. (Corresponding to such a rise in deposits there is also necessarily lower borrowing by the Treasury from the Bank of France and/or net purchases of securities by the financial intermediaries in the sphere of the Treasury, as more balance sheet information would show.) However, these reduced-form coefficients of the influence of i_1 can be questioned and need further corroboration (compare Melitz, 1977, Part II, p. 96).

The other reduced-form coefficients in Table 1 accord with a priori expectations. Their only drawback are many wide differences in the estimates of the influences on F and $FF/\$$ depending on M_4 or M_1. This is true for example, of the effects of P_x, i^*, $DM/\$$, and C_t on F, and the effects of P_x, i^*, and C_t on $FF/\$$ (compare with the last paragraph of Appendix B, concerning the structural estimates). Despite these differences in the results depending on the money measures, three major conclusions emerge from Table 1:

1) Movements of official reserves offset about 70 percent of changes in CFD. The effects of CFD on money are correspondingly

TABLE 1—THE REDUCED-FORM COEFFICIENTS OF INFLUENCE ON F, $FF/\$$, AND M[a]

	M_4			M_1		
	F	$FF/\$$	M	F	$FF/\$$	M
P_x						
$X = 1$	0.18	−0.0051	3.33	0.29	−0.0086	1.60
	(4.0)	(4.0)	(12.7)	(6.3)	(6.3)	(11.5)
$X = 0$	0.18	–	3.44	0.31	–	1.70
	(3.9)		(34.4)	(6.0)		(17.0)
$DM/\$$						
$X = 1$	1.09	0.99	20.4	1.77	0.97	9.84
	(4.0)	(29.5)	(29.5)	(6.2)	(30.2)	(30.2)
$X = 0$	–	–	–	–	–	–
i_1						
$X = 1$	−1.48	0.043	0.88	−2.56	0.076	0.78
	(4.0)	(4.0)	(4.0)	(7.7)	(3.3)	(7.4)
$X = 0$	1.52	–	0	−2.70	–	0
	(3.9)			(7.7)		
i^*						
$X = 1$	−2.25	0.065	1.35	−1.33	0.040	1.35
	(5.5)	(5.3)	(5.3)	(3.3)	(3.3)	(3.3)
$X = 0$	−2.33	–	0	−1.40		0
	(5.5)			(3.4)		
λ						
$X = 1$	0.83	0.024	−0.49	−0.38	0.011	0.12
	(0.74)	(0.74)	(0.74)	(0.4)	(0.4)	(0.4)
$X = 0$	0.86		0	0.40	–	0
	(0.74)			(0.4)		
dP_a						
$X = 1$	5.14	0.15	1.2	4.95	0.15	−6.40
	(8.6)	(8.3)	(0.7)	(7.0)	(6.8)	(4.2)
$X = 0$	5.30	–	1.90	5.22	–	−7.91
	(8.7)		(1.1)	(7.1)		(4.8)
CFD						
$X = 1$	0.71	0.021	0.43	−0.67	0.020	0.20
	(12.4)	(11.5)	(11.5)	(14.9)	(13.0)	(13.1)
$X = 0$	0.74	–	0	−0.70	–	0
	(12.4)			(14.5)		
Y_p						
$X = 1$	0.19	−0.0055	3.12	0.34	−0.010	0.93
	(3.2)	(3.2)	(25.6)	(6.1)	(6.0)	(12.4)
$X = 0$	0.19	–	3.23	0.36	–	1.04
	(3.2)		(26.10)	(5.9)		(12.6)
C_t						
$X = 1$	0.28	−0.0080	−0.16	0.17	−0.0050	−0.051
	(7.3)	(7.0)	(7.0)	(5.0)	(4.9)	(4.9)
$X = 0$	0.28	–	0	0.18	–	0
	(7.5)			(5.2)		

[a]The coefficients of influence refer to effects of changes in one percentage point in P_x, $DM/\$$, i_1, i^*, λ, and dP_a, and one billion francs in CFD, Y_p, and C_t. The terms M_4, M_1, and F are in units of one billion francs. These coefficients concern total cumulative impacts over time. See Appendix B for more information.

moderate, those on M_4 being naturally larger than the ones on M_1. As mentioned earlier, the large offset coefficients of CFD are plausible because virtually all financing of fiscal deficits took place through the central bank rather than bond issues to the private sector in most of the study period. According to these estimates then, any genuine open market operations in France, if they took place, would be offset by 70 percent or somewhat more

(since there would then be no diminution of the offset coefficient through bond financing).

2) The influences on official reserves under fixed exchange rates are only slightly, almost negligibly, smaller under flexible rates. Thus, what used to put downward (upward) pressure on official reserves in France before 1971.4 still causes virtually the same decrease (increase) in official reserves. Despite this, however, the franc was able to undergo all of the substantial movements in Figure 2. If small deviations from the levels of official interventions that would be required to keep the franc fixed can thus bring about large movements in the franc, the conclusion can only be that the stabilizing influence of the franc is weak, at least in the short run, in line with the literature on "overshooting."

3) Since the abandonment of fixed rates, the $DM/\$$ rate has become a dramatic influence on official reserves, the franc, and money in France. This can be seen by comparing the coefficients of $DM/\$$ in Table 1, first, with the mean values of F, M_4, and M_1 in the study period (see Appendix B) and second, with the other coefficients in the table.

In conclusion, this study may be said largely to constitute an attempt to rescue the monetary approach from a certain simplism in theoretical formulation and test procedure which could prove more damaging to it in the long run than any amount of criticism. The outcome, we hope, is a more persuasive confirmation of the value of the approach, despite a clear view of some of its limitations.

Appendix A

A = Net commercial bank assets
B = Reserve base
B_m = Base money
C_b = Loans to the private sector financed by the commercial banks themselves; or commercial bank credit in this sense
C_t = Loans to the private sector financed by the public credit institutions under Treasury control
CFD = The cumulative value of fiscal deficits
D_b = Commercial bank deposits
D_{po} = Postal checking deposits
D_{sb} = Deposits at the national savings banks
$DM/\$$ = Price of the *U.S.* dollar in terms of the deutschmark
F = Official foreign reserves
$FF/\$$ = Price of the *U.S.* dollar in terms of the French franc
i = The domestic interest rate, i.e., the interest rate on bank credit
i^* = The foreign interest rate
i_1 = The central bank rate, i.e., bank rate
M_4 = Money in the sense of currency plus total demand deposits (inclusive of postal checking accounts), and total savings deposits (inclusive of those with the state savings banks as well as with the commercial banks)
N = Currency
P = Price level
P_x = Exogenous component of the price level
dP_a = Anticipated rate of inflation
r = R_r/D_b, weighted-average legal reserve requirement
R = Commercial bank reserves
R_e = Commercial bank excess legal reserves
R_r = Commercial bank legally required reserves
RD = Rediscounts or commercial bank refinancing
TAR = Target value of $FF/\$$ under fixed exchange rates
X = Dummy variables for the system of flexible exchange rates
Y_p = Permanent real income
Y_{tr} = Transitory real income
λ = The competitive legal disadvantage of the commercial banks relative to the national savings banks in the issue of savings deposits
ψ_4 = Competitive position of commercial bank deposits relative to base money

Appendix B: The Full-Information Maximum Likelihood Estimates of the Structural Form

With respect to data, we used quarterly averages for all stocks, interest rates, and foreign exchange rates. We used the three-month Euro-dollar rate for the foreign interest rate i^*, the rate on commercial bank overdrafts for the rate on domestic bank credit i, the legal rate on traditional savings deposits at the savings banks (*les dépôts traditionnels*) for i_d, and the Bank of France's daily rate at the "open market" desk for the cost of refinancing at the central bank, i_1. The additional interest rate expression λ is the excess of the legal rate on the traditional savings deposits at the savings banks over the legal rate on the identical type of deposits at the commercial banks (*les comptes sur livrets*), plus an adjustment for the tax advantage on the interest income received from the savings banks. Our measure of real income Y is the real domestic product. The price level P is the implicit deflator of Y. Regarding permanent income and anticipated inflation, we used the traditional measures:

$$Y_p = [\sum_{i=1}^{19} (\frac{0.9}{1+g})^i]^{-1} \sum_{i=1}^{19} (0.9)^i Y_{t-i};$$

$$dP_a = [\sum_{i=1}^{19} (0.9)^i]^{-1} \sum_{i=1}^{19} (0.9)^i dP_{t-i}$$

The variable g in the formula for permanent income is an adjustment for secular growth, and the 0.9 coefficient in both formulas, which is standard, was found to perform better than lower decimal values in earlier single equation estimates of the demand for credit. With respect to real transitory income Y_{tr}, or $Y - Y_p$, the negative values following the political disturbances of May–June 1968, completely dominate the French series for 1962.1 to 1974.4. Therefore we use a dummy variable for the last three quarters of 1968 as an index of Y_{tr}. Except for the three-month Euro-dollar rate, which comes from the *International Financial Statistics*, our basic sources of statistics are the Conseil National du Crédit and the Institut National de la Statistique et des Etudes Economiques. More information about construction of the series, original sources, and final data inputs, may be obtained directly from the authors.

The full-information maximum likelihood estimates of equations (8)–(13) over 1962.1–1974.4 are as follows for M_4:

$$(8)\quad \frac{C}{P_x + 0.06FF/\$} = \underset{(29)}{-382} - \underset{(4.21)}{8.32i} + \underset{(5.51)}{6.68i^*} + \underset{(5.82)}{13.73\, dP_a} + \underset{(19)}{2.75\, Y_p} - \underset{(2.42)}{5.10\, Y_{tr}}$$

$$(9)\quad C = \underset{(12.6)}{-268} + \underset{(12.4)}{45.24i} - \underset{(1.75)}{3.40i_1} - \underset{(4.58)}{26.84\lambda} + \underset{(1.30)}{122r} + \underset{(13)}{3.58F} + \underset{(12.1)}{2.16CDF} + C_t$$

$$(10)\quad A = \underset{(13)}{-242} + \underset{(12.3)}{40.1i} - \underset{(3.84)}{6.67i_1} - \underset{(3.07)}{16.4\lambda} - \underset{(6.3)}{454r} + \underset{(15.1)}{3.41F} + \underset{(15.3)}{2.28CFD}$$

$$(11)\quad \frac{M}{A} = \underset{(21.5)}{2.42} \frac{1}{1-r} - \underset{(9.1)}{0.147} \frac{i}{1-r} + \underset{(9.24)}{0.087} \frac{i_1}{1-r} + \underset{(1.94)}{0.056} \frac{\lambda}{1-r} - \underset{(5.75)}{0.0055} \frac{Y_p}{1-r}$$

$$(12)\quad \frac{M}{P_x + 0.06FF/\$} = \underset{(28.9)}{-332} - \underset{(1.06)}{1.55dP_a} + \underset{(26.1)}{2.63Y_p}$$

$$(13)\quad \frac{FF}{\$} = \underset{(34)}{3.10X} - \underset{(37)}{0.03FX} + \underset{(30)}{1.02} \frac{DM}{\$} X + TAR$$

In order to interpret the coefficients of these equations, it is necessary to note that all stock and flow values are in billions of francs;

i, i^*, i_1, i_d, and λ are in percentage points (for example, a 7 percent interest rate is 7.0); and r is in units of 100 percentage points (for example, a 7 percent legal reserve requirement is 0.07). (This divergent treatment of r is in keeping with the $1 - r$ notation of equation (11).) Thus, the coefficient of i in equation (9), for example, says that a one percentage point rise in i raises the supply of C by 45.24 billion francs. The coefficient of r in this same equation says that a rise of r of 100 percent raises the supply of C by 122 billion francs. These are all cumulative-impact coefficients. The total time of influence, as can be seen from the text, is three years for equation (8) (except for Y_{tr}), one year for equations (9)-(12) (except for r), and one quarter for equation (13). The mean values of C, A, M_4, M_1, and F are, respectively, 321, 193, 422, 209, and 29. The values in parentheses below the coefficients are ratios of parameter estimates to asymptotic standard errors. (The particular *FIML* procedure supposes that all of the errors are serially uncorrelated but involves correction for the covariance of errors between equations.)

The coefficients of determination, or R^2s, depend on the system as a whole and not the quality of any single equation. For C, the R^2 is 0.990, for i it is 0.950, for A, 0.978, for M, 0.990, and for F, 0.698. The R^2 regarding $FF/\$$ is meaningless for the period as a whole, which covers mostly fixed exchange rates. We have therefore constructed a special R^2 for this variable relating strictly to the period of flexible exchange rates 1971.4-1974.4 (thirteen observations). This we have done by taking one minus the sum of the squared errors during the subperiod divided by thirteen times the variance of $FF/\$$ during the period as a whole. The R^2 is then 0.729. The comparable R^2 for F over the same subperiod is 0.647.

In the case of M_1, the *FIML* estimates of the equations are

$$(8')\quad \frac{C}{P_x + 0.06FF/\$} = \underset{(22)}{-365} + \underset{(0.40)}{1.30i} + \underset{(3.31)}{6.44i^*} + \underset{(2.18)}{7.58dP_a} + \underset{(10.7)}{2.23Y_p} - \underset{(2.48)}{6.13Y_{tr}}$$

$$(9')\quad C = \underset{(12.6)}{-304} + \underset{(12)}{58.8i} - \underset{(2.55)}{6.18i_1} - \underset{(5.11)}{30.26\lambda} - \underset{(0.91)}{80r} + \underset{(13)}{3.57F} + \underset{(11.9)}{2.23CFD} + C_t$$

$$(10')\quad A = \underset{(12.9)}{-270} + \underset{(12.3)}{48.5i} - \underset{(4.96)}{10.57i_1} - \underset{(4.24)}{24.74\lambda} - \underset{(7)}{476r} + \underset{(14.6)}{3.33F} + \underset{(14.3)}{2.24CDF}$$

$$(11')\quad \frac{M}{A} = \underset{(14.5)}{1.344}\frac{1}{1-r} - \underset{(10.8)}{0.162}\frac{i}{1-r} + \underset{(9.01)}{0.078}\frac{i_1}{1-r} + \underset{(3.15)}{0.073}\frac{\lambda}{1-r} - \underset{(5.45)}{0.0043}\frac{Y_p}{1-r}$$

$$(12')\quad \frac{M}{P_x + 0.06FF/\$} = \underset{(12.9)}{-62} - \underset{(4.82)}{6.43dP_a} + \underset{(1.46)}{0.85Yp} - \underset{(1.46)}{4.78i_d}$$

$$(13')\quad \frac{FF}{\$} = \underset{(34)}{3.14X} - \underset{(29.32)}{0.03FX} + \underset{(31)}{1.02}\frac{DM}{\$}X + TAR$$

The R^2s, in the case, are 0.992 for C, 0.979 for i, 0.989 for A, 0.988 for M, and 0.856 for F. For the subperiod of flexible exchange rates, the R^2 for $FF/\$$, calculated as before, is 0.858, and that of F, 0.866. It can be seen that the use of M_1 leads to marked improvements in R^2s for i, F, and $FF/\$$. Also, apart from equations (11′) and (12′), where we must expect differences depending on the money measure, there are notable differences in the structural estimates for the two money measures. In particular, the coefficients of i and dP_a in the demand for credit differ in the two cases, that of i being insignificant and of the wrong sign in the case of M_1 but highly significant and of the right sign in the case of M_4. On the other hand, i_1 performs notably better in the supply equations (9) and (10) for M_1 than M_4.

REFERENCES

P. Artus, "Secteur financier" (du modèle Métric), *Annales de l'INSEE*, Apr./Sept. 1977, 319–52.

Robert E. Baldwin et al., *Trade, Growth, and the Balance of Payments: Essays in Honor of Gottfried Haberler*, Amsterdam 1966.

D. Bean, "International Reserve Flows and Money Market Equilibrium: The Japanese Case," in Jacob Frenkel and Harry Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976

K. Brunner and A. Meltzer, "A Credit Market Theory of the Money Supply and an Explanation of Two Puzzles in U.S. Monetary Policy," *Riv. Int. Sci. Econ. Com.*, May 1966, *13*, 405–32.

——— and ———, "Liquidity Traps for Money, Bank Credit, and Interest Rates," *J Polit. Econ.*, Jan./Feb. 1968, *76*, 1–37.

R. Dornbusch, "The Theory of Flexible Exchange Rate Regimes and Macroeconomic Policy," *Scand. J. Econ.*, No. 2, 1976, *78*, 255–75.

——— and P. Krugman, "Flexible Exchange Rates in the Short Run," *Brookings Papers*, Washington 1976, *3*, 537–75.

André Fourcans, *La Politique de la monnaie: analyse et propositions*, Paris 1976.

Jacob Frenkel and Harry Johnson, *The Monetary Approach to the Balance of Payments*, London 1976.

H. Genberg, "Aspects of the Monetary Approach to Balance-of-Payments Theory: An Empirical Study of Sweden," in Jacob Frenkel and Harry Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

H. G. Johnson, "The Monetary Approach to the Balance-of-Payments Theory," in Jacob Frenkel and Harry Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

P. Kouri and M. Porter, "International Capital Flows and Portfolio Equilibrium," *J. Polit. Econ.*, May/June 1974, *82*, 443–67.

J. Melitz, (1976a) "Inflationary Expectations and the French Demand for Money 1959–70," *Manchester Sch. Econ. Soc. Stud.*, Mar. 1976, *44*, 17–41.

———, (1976b) "Un Modèle Général pour la détermination du stock du monnaie et son application à la France: Part I," *Annales de l'INSEE*, Oct./Dec. 1976, 41–75; "Part II," Jan./Mar. 1977, 63–106.

J. Niehans, "Interest Rates and the Balance of Payments: An Analysis of the Swiss Experience," in Robert Baldwin et al., eds., *Trade, Growth, and the Balance of Payments: Essays in Honor of Gottfried Haberler*, Amsterdam 1966.

J. R. Zecher, "Monetary Equilibrium and International Reserve Flows in Australia," in Jacob Frenkel and Harry Johnson, eds., *The Monetary Approach to the Balance of Payments*, London 1976.

Commissariat Général au Plan, *Modèles monétaires de l'économie française*, Paris 1976.

International Monetary Fund, *Int. Finan. Statist.*, various issues.

Growth and the Long-Run Theory of International Capital Movements

By ROY J. RUFFIN*

Typically, long-run international capital movements have been treated in a one- or two-sector growth model of two large countries or one small country facing a fixed world rate of interest. A number of issues have been discussed: the properties of the steady-state solution (see Koichi Hamada; Hajime Hori and Jerome Stein; Philip Neher); the characteristics of optimal foreign investment (see James Hanson; Takashi Negishi); the approach to the steady-state solution (see Stanley Fischer and Jacob Frenkel, 1974a, b; Yusuke Onitsuka); and the impact of changes in the terms of trade on capital movements and gold flows (see George Borts).

This paper features a two-country version of Robert Solow's one-sector growth model with perfect capital mobility between two large countries.[1] The model, of course, is not new (see Hamada; Negishi). However, the absence of an adequate characterization of this basic model has inhibited the development of the theory. I attempt fresh insights by an alternative presentation of the model and by raising a different set of questions.

The model is exhibited in Sections I and II. Section I displays some short-run properties of the model. Section II begins by deriving the condition for a unique steady-state solution; and presents a useful diagram that allows a simple comparison of the steady-state solution with mobile capital and the solution under portfolio autarky.

Section II also raises the first question: Does perfect capital mobility lower the long-run wage rate in the capital-exporting country compared with portfolio autarky? This is the conclusion of the one-commodity models of G.D.A. MacDougall and Murray Kemp, where the conclusion follows from the assumptions of no growth, one commodity, constant returns to scale, and the law of diminishing returns. With growth, however, capital mobility promotes capital accumulation (due to efficiency gains) and it is no longer clear that removing a bit of the capital stock lowers the steady-state capital stock located in a country. The main result is that the capital exporting country does nevertheless experience a reduction in long-run wages. But I leave open the question, perhaps more significant from the standpoint of political economy, of whether a move from portfolio autarky to *some* capital mobility would lower wages in the long-run for the capital-exporting country. This would happen in the no-growth world of MacDougall and Kemp, but still must be proved for a world of growth.

Section III analyzes the transfer problem. The striking characteristic of my discussion is that in the model there is no short-run transfer problem, as the causes are swept away by assumption. This allows us to concentrate on the long-run aspects themselves. I first examine the traditional question of determining the secondary burdens or blessings of a transfer, and find the latter virtually impossible. I then inquire into the more novel problem of determining the cost of the transfer with capital mobility compared to a world of immobile capital. The main conclusion is that there is a presumption that capital mobility will lower transfer costs for capital-importing countries but raise them for capital-exporting countries.

Section IV gives a brief account of what the model implies about the so-called stages in the balance of payments—a question raised

*University of Houston. I am indebted to my colleague Paul Gregory for stimulating my interest in this subject (see Gregory and Joel Sailors). I thank an unknown referee and George Borts for useful comments; I also profited from the remarks of my colleagues Motoshige Itoh, Peter Mieszkowski, John Rowe, and Sailors, plus the comments of Ronald Jones and Elhanan Helpman.

[1]The small-country assumption is made by Fischer and Frenkel (1974b), Borts, and Onitsuka; the two-sector assumption is made by Hori and Stein, Fischer and Frenkel (1974a, b), and Borts.

by Fischer and Frenkel (1974a, b), Hori and Stein, and Onitsuka. I give a partial answer to this question by showing that when two large countries have a significant amount of outstanding loans between them that this may impart a cyclical component to the trade balance in the approach to the steady-state solution.

I. The Model

Consider two economies, the home and the foreign, producing a single, identical good that can be either consumed or used as capital. The good is produced under competitive full-employment conditions and constant returns to scale. The two productive factors, labor and capital, are substitutable.

Let k and k^* denote the *owned* capital per unit of labor in the home and foreign countries, respectively. The labor supplies L and L^* are assumed to grow at the same constant exponential rate n in both countries. The conception of capital movements is simple: the home country simply locates Z units of its capital stock abroad, so that the capital-labor ratio *located* in the home country is $k - z$, where $z = Z/L$. The capital-labor ratio located in the foreign country is $k^* + bz$, where $b = L/L^*$.

The production function in the home country is denoted by $f(k - z)$, and the production function in the foreign country is $g(k^* + bz)$. Both functions are in per capita terms and measure flow rates of output at a particular moment in time. I assume that these production functions satisfy the usual neoclassical conditions:

$$(1)\quad f' > 0; f'' < 0; f'(0) = \infty; \quad f'(\infty) = 0; f(0) = 0; f(\infty) = \infty$$

Similar conditions hold for the $g(\cdot)$ function.

I assume that capital is perfectly mobile; thus, at every point in time,

$$(2)\qquad f'(k - z) = g'(k^* + bz) = r$$

where r is the equalized rate of return on capital in the world. Equations (2) jointly determine the amount of capital located abroad z, and the rate of return r for any prevailing (k, k^*) pair in the world economy. Conditions (1) imply a unique solution.

To determine per capita incomes, we need only note that income equals national output (per capita) plus or minus the investment income from capital located abroad or the debt service flow for the capital-importing country. The home country earns rz and the foreign country pays rbz (z, of course, can be negative), where r and z are determined by (2) above. Thus

$$(3)\qquad y = f(k - z) + rz$$

$$(4)\qquad y^* = g(k^* + bz) - rbz$$

where y and y^* denote per capita incomes in the home and foreign countries.

I have, so far, just summarized the static theory of capital movements, popularized by MacDougall and Kemp. The dynamic part of the story comes from Solow. I assume that each country saves a constant fraction s or s^* of income. Investment per man in each country is then sy or s^*y^*. The amount of investment required per member of the present labor force to maintain the present capital-labor ratio is, as in Solow, nk or nk^*. Accordingly, the capital accumulation or decumulation equations are simply:

$$(5)\qquad Dk = sy - nk$$

$$(6)\qquad Dk^* = s^*y^* - nk^*$$

where the operator D indicates the time derivative of the variable.

I have now laid out the entire model. Equations (2), (3), and (4) determine y and y^* for any given pair (k, k^*), and equations (5) and (6) determine the course of the economic system through time.

How the model behaves in the long run depends on how k and k^* affect z, r, y, and y^* in the short run. From (2), the effects on foreign investment are

$$(7)\qquad \partial z/\partial k = f''/(f'' + bg'') > 0$$

$$(8)\qquad \partial z/\partial k^* = -g''/(f'' + bg'') < 0$$

Thus an increase in k increases the home country's per capita foreign investments, while an increase in k^* reduces the home

country's per capita foreign investments. Both of these results follow from the law of diminishing returns. On the other hand, an increase in either k or k^* will depress the rate of return on capital; in particular,

$$(9) \qquad \partial r/\partial k = bf''g''/(f'' + bg'')$$
$$\text{and } \partial r/\partial k^* = f''g''/(f'' + bg'')$$

where it is clear that

$$(10) \qquad \partial r/\partial k = b\partial r/\partial k^* < 0$$

Equation (10) is merely the requirement that an increase in the world stock of capital has the same impact on r, regardless of ownership.

Consider now the effects of changes in k or k^* on per capita incomes. Differentiation of y, equation (3), with respect to k yields (noting $f' = r$)

$$(11) \qquad \partial y/\partial k = r + z\partial r/\partial k$$
$$(12) \qquad \partial y/\partial k^* = z(\partial r/\partial k)/b$$

where (12) uses equation (10). Differentiation of y^*, equation (4), yields

$$(13) \qquad \partial y^*/\partial k^* = r - z\partial r/\partial k$$
$$(14) \qquad \partial y^*/\partial k = -bz\partial r/\partial k$$

where, again, (10) is used in (13).

The most important result emerging from the above equations is found in comparing equations (12) and (14): $\partial y/\partial k^*$ and $\partial y^*/\partial k$ have opposite signs if z is nonzero (positive or negative). This important asymmetry arises from the existence of foreign investments or, from the other point of view, foreign debt. Suppose the home country is the creditor country ($z > 0$). Since $\partial r/\partial k < 0$, it follows that $\partial y/\partial k^*$ would be negative and $\partial y^*/\partial k$ would be positive. Intuitively, an expansion in the debtor (creditor) country's owned capital-labor ratio depresses r and, consequently, hurts (benefits) the creditor (debtor) country.[2]

[2]The term "debtor" is only meant to imply the obligation to pay competitive rents on the use of foreign capital rather than fixed interest obligations. Strictly speaking, this paper is only concerned with equity investments.

II. The Steady State

This section investigates the steady-state solution. In particular, we compare the steady-state solutions under autarky (prohibited capital movements) with the solution under free capital mobility.

Under free mobility, the capital accumulation equations are

$$(15) \qquad Dk = s\{f(k - z) + rz\} - nk$$
$$(16) \qquad Dk^* = s^*\{g(k^* + bz) - rbz\} - nk^*$$

Under autarky one would restrict z to equal 0, which would be the Solow model.

The steady state is defined as the pair $(\bar{k}, \bar{k}^*)$ that satisfies $Dk = Dk^* = 0$. The system will be stable if we assume the following conditions:

$$(17) \qquad \partial(Dk)/\partial k = a_{11} = s(r + z\partial r/\partial k) - n < 0$$
$$(18) \qquad \partial(Dk^*)/\partial k^* = a_{22} = s^*(r - z\partial r/\partial k) - n < 0$$

Define also

$$(19) \qquad \partial(Dk)/\partial k^* = a_{12} = sz(\partial r/\partial k)/b$$
$$(20) \qquad \partial(Dk^*)/\partial k = a_{21} = -s^*bz\partial r/\partial k$$

Now, the (local) stability conditions are

$$(21) \qquad a_{11} + a_{22} < 0$$
$$(22) \qquad \Delta = a_{11}a_{22} - a_{12}a_{21} > 0$$

Since (17) and (18) imply (21) and (22), noting $a_{12}a_{21} < 0$, the system is stable. When $z = 0$, the stability condition becomes $n - sr > 0$ (for the home country), which is just the usual stability condition in the Solow model

As we shall see later, conditions (17) and (18) also guarantee a unique steady-state solution. It is important, therefore, to establish a more basic result:

LEMMA: *Condition* (17) *for the home country* (or (18) *for the foreign country*) *will be satisfied in the vicinity of the home* (*foreign*) *country's steady-state solution if and only if the marginal social product of labor in the home* (*foreign*) *country is positive.*

FIGURE 1

To prove this, we need only consider the home country. Total income is Ly; thus the marginal social product of labor is $\partial(Ly)/\partial L = y - k\partial y/\partial k$. Using (11) we see that $\partial(Ly)/\partial L = y - kr - kz\partial r/\partial k$. But in the steady state we must have $y = nk/s$. Thus, a positive marginal social product of labor implies $n > s(r + z\partial r/\partial k)$, which is condition (17).

This is similar to the Solow condition for uniqueness and stability, since $n > sr$ implies the wage rate is positive and the wage rate is the marginal social product of labor in the Solow model. Needless to say, we shall assume the marginal social product of labor is positive throughout this paper.

Now, we are ready to examine the steady state. In Figure 1, the curve $z = 0$ depicts all the combinations of k and k^* consistent with zero capital movements under free capital mobility. It comes from equation (2) with $z = 0$ and is upward sloping, because an increase in k must entail an increase in k^* in order to eliminate the incentive for capital to migrate. (For the moment, ignore the remaining curves.) Southeast of $z = 0$, z will be positive or the home country will be the creditor nation; northwest of $z = 0$, z will be negative or the foreign country will be the creditor nation.

To understand the rest of Figure 1, consider point A. Assume that point A corresponds to the steady-state capital-labor ratios prevailing in the two countries under autarky (for example, for the home country, $sf(k_A) = nk_A$). This implies, if capital movements are allowed, that capital will migrate to the foreign country. Now, define the $Dk = 0$ curve as all the combinations of k and k^* such that the home country is in steady-state equilibrium under free trade. Point B, directly above A on $z = 0$, must be a solution to $Dk = 0$; for when $k^* = k^*_B$, by the uniqueness of the solution to $sf(k_A) = nk_A$, we know that $Dk = 0$ when $k = k_A$. To determine the slope of $Dk = 0$, we know from the implicit function rule that

$$(23)\quad (dk/dk^*)_{Dk=0} = -a_{12}/a_{11} = s(z/b)(\partial r/\partial k)/\{n - s(r + z\partial r/\partial k)\}$$

Since the denominator is positive by the above lemma, and in the numerator $\partial r/\partial k$ is negative, the slope will be positive or negative as z is negative or positive. This proves that $Dk = 0$ is upward sloping above $z = 0$ and downward sloping below $z = 0$. The $Dk = 0$ curve must appear as shown in Figure 1 and cannot cut the $z = 0$ curve again; otherwise, the uniqueness of the autarkic solution (implied by our assumptions) would be violated. By similar reasoning, the slope of the $Dk^* = 0$ curve will be upward sloping *below* the $z = 0$ curve and downward sloping *above* the $z = 0$ curve—the exact opposite—and must appear as shown in Figure 1. For completeness, its slope is

$$(24)\quad (dk/dk^*)_{Dk^*=0} = -a_{22}/a_{21} = -\{n - s^*(r - z\partial r/\partial k)\}/s^*bz\partial r/\partial k$$

In the same region, the two curves have opposite slopes, arising from the sign difference between a_{21} and a_{12} implied by the existence of a nonzero level of per capita foreign investments z.

Since the $Dk = 0$ and $Dk^* = 0$ curves must have the shapes depicted in Figure 1, given our assumption about a positive marginal social product of labor, we know that the steady-state solution is unique.

Figure 1 shows that the capital-labor ratios owned by both countries in the steady-state solution with free capital mobility *both* exceed the capital-labor ratios in the steady-state solution under autarky, since point S must be northeast of point A. An immediate corollary is that per capita incomes are higher with capital mobility, since any steady state requires $sy = nk$ and a higher k requires a higher y to sustain it. Thus:

THEOREM 1: *The steady-state solutions for per capita incomes and the capital-labor ratios with perfect capital mobility exceed the steady-state solutions with prohibited capital movements for both countries.*[3]

This theorem, of course, is perfectly analogous to the usual gains from trade result that free trade is better than no trade. It does not imply that a country could not benefit from a sufficiently small tax on foreign investment income (see Kemp and Negishi).

How are long-run wage rates and interest rates affected by the transition from autarky to free capital mobility? For the foreign country, the capital-importing country, the answer is easy. The foreign country not only owns more capital per unit of labor, but the home country has located some of its capital there as well—so compared to autarky in the foreign country, the rate of interest is lower and the wage rate is higher under free mobility. What happens in the capital-exporting (home) country requires further analysis, because both k and z have increased and it is not clear what happens to the amount of capital per man that is located in the home country.

The question of what happens to home wages and interest rates can be put in different terms. It is clear from Figure 1 that the autarky solution for the home country is the same as the free mobility solution when $k^* = k^*_B$. Thus, we need only inquire into what happens to the home country in the long-run when k^* is *reduced* and held there indefinitely. In general, we know that

$$dr = (\partial r/\partial k)\, dk + (\partial r/\partial k^*) dk^*$$

Since $\partial r/\partial k = b\partial r/\partial k^*$, we can rewrite this as

$$dr = dk^*(dk/dk^* + 1/b)(\partial r/\partial k)$$

From (23) we can calculate dk/dk^* when $Dk = 0$; accordingly, substituting (23) into the above expression and simplifying, we find

$$(25)\quad (dr/dk^*)_{Dk=0} = -(\partial r/\partial k)(n - sr)/ b(sr - n + sz\,\partial r/\partial k) < 0$$

which is negative since the numerator is positive and the denominator, by (17), is negative. This establishes that if we lower k^*, we raise r in the home country or lower wages in the long run. Thus, we have established the result:

THEOREM 2: *Perfect capital mobility lowers long-run wages and raises long-run interest rates in the capital-exporting country compared to autarky; the opposite holds in the capital-importing country.*

An alternative way of establishing that r and k^* are negatively related along the downward-sloping section of the $Dk = 0$ curve (which is obvious along the upward-sloping section) can be seen by considering the following simple experiment. Suppose k^* is at some level where $Dk = 0$ and z is positive. Now reduce k^* and increase k by just enough to keep the rate of interest constant. This implies that $k - z$ is constant; for, otherwise, r would change. Thus, $dk = dz$. The change in per capita income must be rdk, and the change in investment per capita must be $srdk$. But the change in investment per capita required to keep k constant is ndk. Since $n > sr$, it follows that capital will decumulate in the home country, which drives r up and wage rate down from the initial position.[4]

[3]Hamada obtains the result that the world capital-labor ratio is increased in moving from autarky to free capital mobility, but he assumes production functions are identical and saving rates different between countries. My result is more general.

[4]The above proof also shows that our assumption about constant saving rates is not crucial; for we could have assumed that the saving rates respond to the rate of interest and still the reasoning in this paragraph would be unaltered.

Aside from the obvious policy implications, the above theorem is interesting because it demonstrates that the static effects of capital mobility on wages are stronger than the dynamic effects. Reconsider Figure 1. Starting at autarky, point A, when capital mobility is allowed, there will be an immediate reduction in wages in the home country (static effect). When world capital accumulates, moving world economy to point S, the rate of interest must fall and wages must rise. But the import of Theorem 2 is that the dynamic increase in wages associated with the induced capital accumulation is less than the initial reduction in wages in the home country that started the whole process.

III. The Transfer Problem

The crucial question in transfer theory is whether the transfer confers a secondary burden or secondary blessing on the paying country. In the traditional model, which abstracts from both growth and capital mobility, all depends on the sum of the marginal propensities to import: if the sum is less than unity a secondary burden is carried and if the sum is greater than unity a secondary blessing is enjoyed (see Paul Samuelson and Ronald Jones). The orthodox presumption is that there is a secondary burden. In this section I show that the orthodox position is strongly supported in the present model. In addition, it will be shown that there is a presumption that capital mobility raises transfer costs for capital-exporting countries and lowers them for importing countries. This latter presumption follows from Theorem 2.

A transfer should be conceived of as a permanent flow that is being transferred between countries. Should a lump sum payment be involved, the transfer out of income is the stream of payments equivalent to the lump sum amount. We shall here assume that the transfer consists of a constant flow of goods per capita sent from one country to the other in perpetuity.

The "short-run" or static effects of a transfer in the present model are clear: there is no transfer problem. Since there is only one good in the world, transferring the good from one country to the next creates an immediate export equal to the transfer. In effect, the paying country has a marginal propensity to import of zero and the receiving country has a marginal propensity to import of unity out of the transfer. Moreover, for any given world distribution of capital-labor ratios, (k, k^*), the per capita amount of foreign investments z will still be the same regardless of the quantity of the transfer. Thus, the equation, repeated here for convenience, still holds: $f'(k - z) = g'(k^* + bz) = r$.

The transfer, however, does influence the amount of saving in each country. We assume that the saving rate is a constant fraction of income net of or including the transfer. Accordingly, the capital accumulation equations become

$$(26) \quad Dk = s\{f(k - z) + rz - t\} - nk$$

$$(27) \quad Dk^* = s^*\{g(k^* + bz) - rbz + bt\} - nk^*$$

where t is the per capita transfer from the home country to the foreign country (bt is the transfer in terms of foreign labor). Throughout this section we shall assume t is positive or that the home country is the paying country.

What is the steady-state impact of t? Setting $Dk = Dk^* = 0$ and differentiating (26) and (27) with respect to t, we obtain

$$(28) \quad d\bar{k}/dt = -s(n - s^*\bar{r})/\Delta < 0$$

$$(29) \quad d\bar{k}^*/dt = s^*b(n - s\bar{r})/\Delta > 0$$

where a bar over a variable indicates a steady-state solution and Δ is defined as in equation (22). For our present purposes, we need to examine Δ more closely. Substituting (17)–(20) into (22) and simplifying yields

$$(30) \quad \Delta = (n - s\bar{r})(n - s^*\bar{r}) + n\bar{z}(s^* - s)\partial r/\partial k$$

We can also calculate the impact of t on the world rate of interest r. This turns out to be, recalling equation (10),

$$(31) \quad d\bar{r}/dt = (\partial r/\partial k)\{d\bar{k}/dt + (1/b)d\bar{k}^*/dt\} = n(s^* - s)(\partial r/\partial k)/\Delta$$

Since the home country is the paying country,

the transfer raises (lowers) world r as the home country's saving rate exceeds (is less than) the foreign country's saving rate. If the saving rates are equal, the transfer does not influence the long-run interest rate.

The change in long-run per capita income induced by the transfer is

$$d\bar{y}/dt = \bar{r}d\bar{k}/dt + \bar{z}d\bar{r}/dt \tag{32}$$

which follows from differentiating $y = f(k - z) + rz$ and using equation (2). The first term in (32), $\bar{r}d\bar{k}/dt$, must be negative: when the home country transfers goods to the foreign country, in the long run it must suffer a loss of capital (equation (28)) with the cost of this loss being evaluated at the rate of interest. The second term in (32) can be positive or negative, depending on what happens to the rate of interest and whether the home country is a capital exporter. This is the "terms of trade" effect on the capital account. If the home country, for example, exports capital and $s > s^*$, the rate of return will rise (equation (31)) and the terms of trade on the capital account will improve.

Substituting (28) and (31) into (32), we can see that a necessary and sufficient condition for $d\bar{y}/dt$ to be *negative* is

$$n\bar{z}(s^* - s)\partial r/\partial k < s\bar{r}(n \quad s^*\bar{r}) \tag{33}$$

The left-hand side of (33) reflects the gain in per capita income (or loss if negative) due to the change in the terms of trade on the capital account and the right-hand side reflects the cost due to the loss of capital induced by transfer. Notice that right-hand side is always positive.

The only possibility of a secondary blessing is that the terms of trade on capital account improve. We can establish the following result:

THEOREM 3: *If the saving rate of the capital-exporting country is less (greater) than the saving rate of capital-importing country, the paying country experiences deterioration (improvement) in its terms of trade on the capital account.*

This result is easy to see. If the paying country is the capital-exporting country, a relatively low rate of saving implies that a transfer will lower the world rate of interest and thus reduce interest receipts from foreign investments; and if the paying country is the capital-importing country, a relatively high rate of saving implies that a transfer will raise the world rate of interest and thus increase the debt burden.

If the saving rate of the capital-exporting country exceeds the saving rate of the capital-importing country, the terms of trade on the capital account will improve for the paying country. Can this terms of trade effect ever be as large as the cost of the induced loss in the capital stock? A realistic appraisal of condition (33) indicates that it is highly improbable, since even in the most extreme cases the former is likely to be a fraction of the latter.[5] The orthodox presumption of a secondary burden is more difficult to escape in the present model than in the standard trade model.

Now, consider the question of whether capital mobility raises the cost of the transfer. Let us consider the comparative steady-state costs of the transfer in two situations: 1) when capital is completely mobile; and 2) when capital is completely immobile. To simplify the comparison, note that in any steady state we must have $s(y - t) = nk$. Hence, $dy/dt = (n/s)dk/dt + 1$. Thus, since n/s and unity are constants, different steady-state transfer costs, (dy/dt) can be compared by merely comparing values of dk/dt.

In the case of autarky, we must have $s\{f(k_A) - t\} = nk_A$. Thus

$$dk_A/dt = -s/(n - sr_A) \tag{34}$$

[5]Consider the following example: $b = 1$, $f(k - z) = 77.679961(k - z)^{.3}$, $g(k^* + z) = 685.95758(k^* + z)^{.1}$, $n = .04$, $s = .25874126$, and $s^* = .00294985$. Then in the steady-state equilibrium $r = .06$, $k = 7400$, $k^* = 100$, and $z = 2400$. This means that 96 percent of the capital located in the foreign country is owned by residents in the home country. A (per capita) dollar transferred in perpetuity from the home country to the foreign country will reduce home income by \$.418, the loss due to the cost of capital effect is \$.55 and the gain due to the terms of trade effect is \$.132 so that the former exceeds the latter by over four times even in this extreme case. This example indicates that with Cobb-Douglas production functions a secondary blessing may be impossible.

where r_A is the rate of interest in autarkic steady state.

Now, compare $d\bar{k}/dt$, from equation (28), with (34). Capital mobility will raise the cost of the transfer if, and only if, $-d\bar{k}/dt$ exceeds $-dk_A/dt$. Using (30) and dividing the numerator and denominator of (28) by $(n - s^*\bar{r})$, we have

$$s/[(n - s\bar{r}) + nz(s^* - s)(\partial r/\partial k)/(n - s^*\bar{r})] > s/(n - sr_A)$$

But, we need only compare the denominators; accordingly, we obtain the condition for capital mobility to raise transfer costs (necessary and sufficient):

$$(35) \quad n(s^* - s)\bar{z}\partial r/\partial k < s(\bar{r} - r_A)(n - s^*\bar{r})$$

If the left-hand side of (35) *exceeds* the right-hand side, capital mobility *lowers* the cost of the transfer compared to portfolio autarky.

Notice this condition is almost the same as (33); the only difference is that the right-hand side of (35) has $-sr_A(n - s^*\bar{r})$ subtracted from the right-hand side of (33). The interpretation of this inequality is straightforward. We are comparing the cost of the transfer under autarky with that of free capital mobility. Under autarky, there are no terms of trade effects. The left-hand side, then, captures the terms of trade gains from having capital mobility (which can be negative). Each term on the right-hand side measures the cost of the induced capital decumulation from the transfer, the opportunity cost of which is $\bar{r}$ with mobility and r_A with capital immobility. Hence, the right-hand side captures the excess of the cost with mobility over this cost with immobility.

Theorem 2 above implies that the right-hand side of (35) will be positive if the paying (home) country is a capital exporter ($\bar{r} > r_A$) and negative if the paying country is a capital importer ($\bar{r} < r_A$). If we assume that $s \leqq s^*$, the left-hand side will be negative or positive as the paying country is a capital exporter ($\bar{z}\partial r/\partial k < 0$) or a capital importer ($\bar{z}\partial r/\partial k > 0$). Accordingly, we have established:

THEOREM 4: *If the paying country has a saving rate no higher than the receiving country's saving rate, capital mobility will raise (lower) the cost of the transfer when the paying country is a capital exporter (importer).*

If the paying country's saving rate exceeds the receiving country's saving rate ($s > s^*$), there is an apparent ambiguity. When the paying country is a capital exporter, $s > s^*$ implies r will rise with a transfer or the terms of trade on the capital account will improve for the paying country. This will partially offset the higher opportunity costs of the sacrificed capital with capital mobility over the case of capital immobility. Thus, in principle, it is possible for capital mobility to lower the cost of the transfer for the capital-exporting country. Similarly, when the paying country is a capital importer, $s > s^*$ implies the terms of trade on the capital account will deteriorate. Thus, in principle, capital mobility could raise the cost of the transfer for a capital-importing country. However, these cases are unlikely due to the relatively small size of the terms of trade effects.[6]

It is clear from Theorem 4 that there is a presumption that capital mobility will raise (lower) the transfer costs of capital-exporting (importing) countries. This suggests, first, that restrictions on capital mobility have a differential impact on the transfer burdens of capital-exporting and importing countries. Such burdens count as part of the potential gains and losses of restrictions on capital movements. Second, the result *could* mean that capital-exporting countries have higher

[6]In the example discussed in fn. 5, the secondary burden of a per capita dollar transferred in perpetuity is \$.429 without capital mobility compared to \$.418 with capital mobility. Thus, in that extreme example, capital mobility lowers the secondary burden of a transfer for the capital-exporting country. However, if the example is changed slightly so that instead of 96 percent of foreign capital being owned by the home country we have 80 percent of the capital owned by the home country (which is still an extreme case), the results are reversed. With $s = .25$ and $s^* = .0145$, we have $z = 2000$, $k = 7000$, and $k^* = 500$ and the secondary burden without capital mobility is \$.429 (unchanged) and with capital mobility is \$.433. This is the normal expectation.

transfer burdens than capital-importing countries, provided under portfolio autarky the transfer burdens were similar. This latter result has practical implications, for example, for evaluating the cost of foreign aid to different countries.

IV. Stages in the Balance of Payments

Since the work of J.E. Cairnes economists have been intrigued about the possibility of stages in the balance of payments (see Fischer and Frenkel, 1974b). The thesis is that a mature debtor or creditor country will have quite a different balance of trade than a young debtor or creditor country. What happens is that eventually interest payments on old loans will exceed fresh loans and so the debtor country will switch from having a trade deficit to a trade surplus. This, of course, is not much more than arithmetic.

Let us first look at Cairnes' arithmetic, using my model and notation, and then turn to the economics of the matter. Recall that per capita foreign investments by the home country is defined by $z = Z/L$, where Z is the total stock of capital invested in the foreign country. The per capita capital outflows from the home country are measured by DZ/L. This can be decomposed into

$$DZ/L = Dz + nz \tag{36}$$

where again D is the operator indicating a time derivative. The first component of (36), Dz, can be called "transitory (per capita) capital outflow" and the second component, nz, is the "steady-state (per capita) capital outflow." If $Dz = 0$, nz is the amount of capital outflow per capita that is required to keep z constant.

The per capita trade balance—denoted here by x—is (by defintion) the per capita outflow of capital DZ/L, minus per capita interest receipts rz. Thus, using (36),

$$x = Dz + (n - r)z \tag{37}$$

In the steady state, $Dz = 0$ and, hence, $x = (n - r)z$. If the home country is a net creditor, $z > 0$, the mature creditor country will have a trade deficit provided $r > n$, just as in Cairnes. The condition for Cairnes' arithmetic to hold ($r > n$) is a well-known necessary condition for dynamic efficiency (see Edmund Phelps). Presumably, this is the main economic content of Cairnes' thesis.

The interesting question, initially raised by Fischer and Frenkel (1974b), is: What is the behavior of the trade balance as the economy approaches the steady state? Can x be expected to behave cyclically? Fischer and Frenkel (1974b) focused on Dz rather than x. If Dz behaves cyclically, then we can at least say that the trade balance x has a cyclical component.

It is easy to show that Dz will behave cyclically, that is, oscillate between positive and negative values, if the adjustment of the owned capital-labor ratios (k, k^*) to the steady state is cyclical. This is because

$$Dz = (\partial z/\partial k)Dk + (\partial z/\partial k^*)Dk^* \tag{38}$$

where it will be recalled that $\partial z/\partial k > 0$ and $\partial z/\partial k^* < 0$ (as in equations (7) and (8)). When capital is accumulating in the home country and decumulating in the foreign country, Dz will be positive; and when capital is decumulating in the home country and accumulating in the foreign country, Dz will be negative.

Thus we must study the solution to the differential equations (15) and (16) describing the capital accumulation paths of k and k^*. The system will exhibit a cyclical adjustment (mathematically, a spiral point) if the roots of the equation

$$\lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0$$

are complex with negative real parts. This will be the case if

$$(a_{11} - a_{22})^2 + 4a_{12}a_{21} < 0$$

Substitution of (17) through (20) into the above inequality yields the condition

$$(s - s^*)^2\{r^2 + (z\partial r/\partial k)^2\} < -2rz(\partial r/\partial k)(s^2 - s^{*2}) \tag{39}$$

Notice that if $z > 0$ and $s^* > s$, this condition cannot be satisfied. A necessary condition, then, is that $s > s^*$ when $z > 0$. This is eminently reasonable, because it requires the capital-exporting country to have the highest

saving rate.

An example of a reasonable case in which a cyclical adjustment occurs can be easily constructed.[7] But the economic possibility can be seen by recalling a proposition discovered in Section I above: an increase in the debtor country's (owned) capital-labor ratio hurts the creditor country through depressing the latter country's earnings on capital account; and an increase in the creditor country's (owned) capital-labor ratio helps the debtor country through depressing the latter country's debt service flow.

Keeping this relationship in mind, imagine that the creditor country is in the steady state (momentarily), but that the debtor country is on an accumulation path. As the debtor country grows, this hurts the creditor country and sends it on a decumulation path. But this decumulation path in the creditor country hurts the debtor country and so induces a retardation of that country's growth path until the debtor country reaches a steady state. If the creditor country is still decumulating (which is what is assured by condition (39)) at that point, the debtor country will start to deculumate capital. This benefits the creditor country and contributes to bringing its decumulation path to an end and, eventually, sending it into an accumulation path. This is exactly the reverse of the beginning of the story and the story keeps repeating until the steady-state solution is reached.[8]

Going back to the condition for the above story, (39), we see that the condition would not be satisfied if $z\partial r/\partial k = 0$. This will hold under two circumstances: (i) if the steady-state equilibrium involves $z = 0$ or (ii) the home country is so small that it exerts no impact on the world rate of interest, i.e., $\partial r/\partial k = 0$. Under either of these circumstances, a cyclical adjustment of Dz is ruled out.

These results partially explain the difficulties Fischer and Frenkel encountered in accounting for the cyclical behavior of Dz. In one paper (1974b) they made the small-country assumption; and, in another (1974a), they assumed two large countries that were identical in all respects so that the steady-state level of z equaled zero

V. Conclusions

This paper has constructed a two-country version of Solow's growth model with perfect capital mobility in such a way as to reveal the relationship between the steady-state solutions under portfolio autarky and perfect capital mobility. It was shown that 1) free capital mobility must lower long-run real wages of the capital-exporting country compared to portfolio autarky; 2) there is a very strong presumption that the orthodox position of secondary transfer burdens holds in such a world; 3) there is a presumption that perfect capital mobility raises (lowers) transfer costs for capital-exporting (importing) countries compared to portfolio autarky; and 4) it is possible for the model to generate cycles in an important component of a country's trade balance in the approach to the steady state.

This paper has made a number of simplifying assumptions. The most important, I believe, are that the analysis has assumed only one commodity and perfect capital mobility. It would be interesting to extend the analysis to the cases involving imperfect capital mobility and several goods.

When commodity trade is considered, capital mobility increases the complexity of the model. Jones and I found, for example, that there may be numerous specialization patterns. However, the basic technique of this

[7] Consider the following example: $b = 1$, $f(k - z) = 142.70485(k - z)^{.25}$, $g(k^* + z) = 77.679961(k^* + z)^{.3}$, $s = .18589744$, $s^* = .17647059$, and $n = .04$. In this case, the solution is $k - z = k^* + z = 5000$, $z = 800$, and $r = .06$; condition (39) is satisfied by a substantial margin. The approach to the steady state must be cyclical. In this paper the term $z\partial r/\partial k$ has played a critical role in some of the conditions (for example, (33), (35), and (39)). The reader may be interested in the fact that for this particular example, which is not an extreme case, $z\partial r/\partial k = -.003472$, which is an extremely small number and accounts for some of my earlier judgements about the size of the terms of trade effects on the capital account.

[8] There are other reasons why capital flows might fluctuate between outflows and inflows. For example, Jones and I show that along the world production-possibility curve in a world with mobile capital, the optimal allocation of the world's capital stock entails a nonmonotonic relationship between the commodity terms of trade and the amounts of foreign investments.

paper may remain viable in such models since for any given capital-labor ratio in one country there ought to be a steady-state capital-labor ratio in the other country even if there are two goods.

REFERENCES

G. Borts, "A Theory of Long-Run International Capital Movements," *J. Polit. Econ.*, Aug. 1964, *72*, 341–59.

John E. Cairnes, *Some Leading Principles of Political Economy*, New York 1874.

S. Fischer and J. Frenkel, (1974a) "Interest Rate Equalization and Patterns of Production, Trade and Consumption in a Two-Country Model," *Econ. Rec.*, Dec. 1974, *50*, 555–80.

——— **and** ———, (1974b) "Economic Growth and Stages of the Balance of Payments," in George Horwich and Paul Samuelson, eds., *Trade, Stability and Macroeconomics*, New York 1974.

P. Gregory and J. Sailors, "Secular Capital Flows Among Industrialized Countries, 1870–1955," unpublished manuscript, Univ. Houston 1978.

K. Hamada, "Economic Growth and Long-Run Capital Movements," *Yale Econ. Essays*, Spring 1966, *6*, 49–96.

J. A. Hanson, "Optimal International Borrowing and Lending," *Amer. Econ. Rev*, Sept. 1974, *64*, 716–30.

H. Hori and J. L. Stein, "International Growth with Free Trade in Equities and Goods," *Int. Econ. Rev.*, Feb. 1977, *18*, 83–100.

R. W. Jones, "The Transfer Problem Revisited," *Economica*, May 1970, *37*, 178–84.

——— **and R. J. Ruffin,** "Trade Patterns with Capital Mobility," in Michael Parkin and A. R. Nobay, eds., *Current Economic Problems*, Cambridge 1975.

M. Kemp, "Foreign Investment and the National Advantage," *Econ. Rec.*, Mar. 1961, *28*, 56-62.

G. D. A. MacDougall, "The Benefits and Costs of Private Investment from Abroad: A Theoretical Approach," *Econ. Rec.*, Mar. 1960, *27*, 13-35.

T. Negishi, "Foreign Investment and the Long-Run National Advantage," *Econ Rec.*, Dec. 1975, *31*, 738–32.

P. A. Neher, "International Capital Movements Along Growth Paths," *Econ. Rec*, Sept. 1970, *46*, 393--401.

Y. Onitsuka, "International Capital Movements and the Patterns of Economic Growth," *Amer. Econ. Rev.*, Mar. 1974, *64*, 24–36.

Edmund S. Phelps, *Golden Rules of Economic Growth*, New York 1966.

P. A. Samuelson, "The Transfer Problem and Transport Cost," *Econ. J.*, June 1952, *62*, 278-304.

R. Solow, "A Contribution to the Theory of Economic Growth," *Quart. J. Econ.*, Feb. 1956, *70*, 65--94.

A Dynamic Disequilibrium Comparison of Fixed and Free Exchange-Rate Regimes

By Kenneth S. Chan*

For the last twenty years economists have debated the advantages of free and fixed exchange-rate regimes. Milton Friedman argues that if internal prices and wages were inflexible, it would be preferable to allow adjustment to occur through a depreciation of domestic currency. Svend Laursen and Lloyd Metzler, Egon Sohmen, and Murray Kemp argue in favor of free (floating) exchange rates by the familiar insulation properties of free rates. Jerome Stein classifies a conflict (compatible) economy as one in which a decline in output is accompanied by an excess demand (supply) of foreign exchange. When output falls for a compatible economy in a free (fixed) exchange-rate regime, the resultant appreciation of domestic currency (increase in the level of money balances) tends to reinforce (mitigate) the initial decrease in output. When output falls for a conflict economy in a free (fixed) exchange-rate regime, the resultant depreciation of domestic currency (decrease in the level of money balances) tends to mitigate (reinforce) the initial decline in output Stein then concludes that a free (fixed) exchange-rate regime is optimal for the conflict (compatible) economy.

There are two major shortcomings of previous comparisons of different exchange-rate regimes: the first is the lack of disequilibrium behavior. In this paper, I develop a disequilibrium model for the analysis.[1] It will be assumed that the money wage adjusts slowly and transactions can occur at labor market disequilibrium. Unemployment generated from this type of economic behavior is typically involuntary.

The second shortcoming is that the results are limited to static short-run comparisons. Most of the previous analyses are based on the standard Keynesian variable income model with rigid wages and prices.[2] Though wages and prices may be considered as fixed in the short run, they must adjust in the long run. In the literature, these long-run aspects have never been satisfactorily analyzed.[3] Indeed, this leaves a good part of the problem out of the picture. In order to evaluate the overall efficiency of exchange-rate regimes, in addition to short-run comparisons, we should also consider the shapes (or speeds of adjustment) of long-run time paths of different exchange-rate regimes. To overcome these setbacks, I construct a disequilibrium model that traces out the long-run time path of different exchange-rate regimes. Not only will the short-run comparative statics be considered, but also the long-run adjustments of those "sticky" prices.

My analysis shows that Stein's classification can be extended to a long-run dynamic framework. For a conflict (compatible) economy, a free (fixed) exchange-rate regime is superior to a fixed (free) exchange-rate regime, even though the latter regime may have a faster speed of adjustment than the former.

In Section I, the analytical framework of the model is developed. Section II analyzes the short-run level of unemployment for each exchange-rate regime; and Section III is an examination of the long-run time paths for

*Assistant professor of economics, Saint Mary's University. This paper is based on parts of my doctoral dissertation. I am indebted to George Borts, James Hanson, Hiroshi Ono, William Der, Arun Mukhopadhyay, and an anonymous referee for helpful comments I alone bear responsibility for errors.

[1]The body of literature starts from Robert Clower, Don Patinkin, Axel Leijonhufvud, Robert Barro, Herschel Grossman, and Edmond Malinvaud

[2]Previous work by Stein, Kemp, Robert Mundell, and Sohmen are based on the Keynesian $C + I + G$ framework of analysis, more recent works by Robert Lucas are based on the disequilibrium analysis of Barro and Grossman.

[3]Earlier, Dagobert Brito and J. David Richardson did attempt a "disequilibrium-dynamic" approach for the analysis of devaluation.

each regime. In Section IV, a dynamic evaluation of exchange rate regimes is offered plus a discussion of the welfare implications of unemployment.

I. The Basic Model

Let us consider a simplified analytical framework which deals with five economic goods: a nontraded (home) good, an import good, an export good, money, and labor services. There are four markets: the home good market, the foreign exchange market, the money market, and the labor market. There is no investment demand and no international capital flows.[4]

A small-country model is employed for the analysis. Assume that there is only a negligible amount of exports (imports) consumed (produced) at home. Labor is homogeneous and is the only freely mobile factor of production between the two industries.

Let us also assume a two-stage adjustment process. In the first stage (the common notion of short run), a short time interval is used in which the price of home goods adjusts to clear the home good market, and either the exchange rate or money stock[5] in the respective free or fixed exchange-rate regime adjusts to clear the balance of payments. In the second stage (the common notion of long run) I use a long time horizon during which the nominal wage is allowed to adjust to clear the labor market. In other words, the Stage I adjustment for the home good market and the foreign exchange market can be approximated by a Walrasian auctioneering process; that is, allowing recontracting *before* short-run equilibrium, and actual quantity transactions occur only *at* short-run equilibrium. Nominal wage is regarded as a shifting parameter in Stage I adjustments.[6] Prior to the Stage II adjustment, transactions do occur at non-market-clearing wages. Transactions at labor market disequilibrium imply the quantities transacted will be determined on the short side of the market. This quantity constraint, resulting from transaction at labor market disequilibrium, will affect the formulation of firms' and households' behavior in other markets.

The model employs the following notation:

W = Money wage in home currency units

P_H = Money price of nontraded goods (home goods) in home currency

$p_h = P_H/W$ = the domestic price of home goods in labor units

$P_X^*(P_M^*)$ = Money price of export (import) goods in foreign currency

ρ = the exchange rate

$e = \rho/W$ = the price of foreign currency in domestic wage-good units

S_x, S_h, S_m = the supply of exports, home goods, and imports, respectively

$C_m(C_h)$ = the demand for imports (home goods)

L = the excess demand of labor market

H = the excess demand for nontraded (or home) goods

$-B$ = the excess demand for foreign exchange. Thus, B is the definition of the balance of payments in units of foreign currency

Π = the flow demand for money in real terms (deflated by nominal wage)

$M^d(M^s)$ = the stock demand (supply) for money; while $m = M^s/W$

$\bar{L}$ = the total supply of labor and is assumed constant

$l_h^d(l_x^d)$ = the demand for labor from the

[4]For a small country, interest rates are fixed in world markets. Hence investment demand and capital flow are generally fixed and ignored in our analysis.

[5]For simplicity, assume that money stock adjusts very "quickly" in a fixed exchange-rate regime. Though the flow of money may eliminate the balance-of-payments disequilibrium gradually over time, my assumption is still plausible if it is assumed a responsible government will obey the rules of the game and adjust the money supply instantaneously, as is required to maintain balance-of-payments equilibrium.

[6]This approach is somewhat akin in spirit to the "asset approach" or "monetary approach" in international finance (see Rudiger Dornbusch; Michael Connolly and Dean Taylor).

nontraded goods (export goods) sector

Y = the money income, $y = Y/W$ is in real terms. Thus $z = y - \Pi$ is defined as the expenditures for C_h and C_m in real terms

k = the exogenous disturbance parameter. It can be anything including P_X^* and/or P_M^* and/or changes in the desire to hoard, etc.

u = the level of unemployment; u^i (u^r) is the level of unemployment of the fixed (free) exchange-rate regime

$\lambda^i(\lambda^r)$ = the speed of adjustment in the fixed (free) exchange-rate regime during Stage II adjustments.

Note that in the notation the superscripts i, r, and f indicate the corresponding variable belongs to the fixed exchange-rate regime, the free exchange-rate regime, and variables at long-run equilibrium, respectively.

Let us now examine the behavior of firms and households and the four markets when the economy is in the region of excess supply of labor.

A. *The Behavior of Firms*

The model has two sectors; one produces nontradable goods and the other produces export goods which are not consumed at home. Each sector uses a fixed factor, which is not transferable, and labor. Due to my assumption of flexible prices and a labor surplus at the current market wage rate, firms are not constrained in the labor or product markets.[7] A "representative" firm in the export sector maximizes profit R:

$$R = \rho P_X^* S_x - W l_x^d$$

subject to $S_x = S_x(l_x^d)$, $S_x' > 0$ and $S_x'' < 0$ which yields the labor demand and commodity supply equations of the following form:

$$(1) \qquad l_x^d = l_x^d(\underset{(+)}{e \cdot P_X^*}) \text{ and } S_x = S_x(\underset{(+)}{e \cdot P_X^*})$$

The positive sign beneath $e \cdot P_X^*$ stands for the sign of the derivative of S_x with respect to $e \cdot P_X^*$.

Similarly, the labor demand and commodity supply equations in the home good sector are

$$(2) \qquad l_h^d = l_h^d(\underset{(+)}{p_h}) \text{ and } S_h = S_h(\underset{(+)}{p_h})$$

B. *Households*

Since we assume flexible prices, households are never constrained on the commodity markets by a shortage of consumption goods.[8] The representative household perceives a demand-imposed constraint upon its employment[9] which in turn limits or constrains "effective" income (as opposed to income at the full-employment level or "notional" income):

$$(3) \quad y = (\rho P_X^*/W)S_x[l_x^d(\rho P_X^*/W)] + (P_H/W)S_h[l_h^d(P_H/W)] = y(\underset{(+)}{p_h}, \underset{(+)}{e}, \underset{(+)}{P_X^*})$$

Households will allocate this disposable effective real income y between home goods, imports, and the accumulation of money balances.

Assume that the stock demand for money depends on the level of nominal disposable income Y only;[10] i.e.,

[7]Note that in the model, firms are never constrained by the lack of sale in the output market, since 1) in the home good sector, the price of home goods is assumed to adjust rapidly to clear the market, and hence approximated by a Walrasian recontracting process and 2) in the export sector, the demand for exports is infinitely elastic.

[8]Shortages or surpluses of consumption goods cannot exist under our assumptions since the price of home goods adjusts rapidly to clear the market and the supply of imports is infinitely elastic.

[9]Note that households are constrained in their labor supply to the potentially limitless export industry by the nontransferable (or fixed) factor of production in the export sector. Thus, unemployment can still exist even in a small-country model with infinitely elastic export demand.

[10]In other words, M^d does not depend on the price of home goods and imports directly. The explanation for the zero price effect in the existing literature is that households regard all changes in prices as permanent, and that the utility function relating present and future consumption is homothetic. Hence, there is no intertemporal effect and no money illusion. For a detailed discussion, see Kemp (pp. 277–81) and the refences cited there.

$$(4) \qquad M^d = M^d(Y)$$

Making use of the linear homogeneity property of the stock demand for money, we can write a flow demand of money equation Π in labor units:

$$(4') \qquad \Pi = \Pi\left(\frac{M^d}{W} - \frac{M^s}{W}\right) = \Pi(y, m)$$

Effective real income y must equal real flow demand for money plus effective real expenditures z (a standard income-expenditure identity):

$$(5) \qquad y = \Pi + z$$

Note that z is in labor units and is homogeneous of degree zero in money prices, wage, and money stock. The representative household will allocate expenditures amongst home goods and import consumption as follows:

$$(6) \quad C_h = C_h(P_H, \rho P^*_M, z \cdot W) = C_h(p_h, eP_M{}^*, z)$$

$$C_m = C_m(P_H, \rho P^*_M, z \cdot W) = C_m(p_h, eP^*_M, z)$$

Since each demand function is homogeneous of degree zero in money prices and nominal expenditures $z \cdot W$, it may be rewritten in the form of relative prices and real expenditures.

C. *The Home Good Market*

The demand for home goods is based upon effective real expenditures z. On the supply side, there is surplus labor at the market wage rate. Hence, the excess demand function for home goods is

$$(7) \quad H = C_h(p_h, eP^*_M, z) - S_h[l^d_h(p_h)] = H(p_h, e, m, k)$$

where k is the exogenous (shifting) variable that represents the most general form of disturbances, including P^*_X and/or P^*_M.

D. *The Foreign Exchange Market*

In a small-country model the demand for exports and the supply of imports are infinitely elastic. Similar to the home goods market, import demand is formulated in terms of effective income. The supply of exports is formulated under the assumption that there is surplus labor at the current market wage rate. The balance-of-payments equation B is as follows:

$$(8) \quad B = P^*_X S_x - P^*_M C_m = P^*_X S_x[l^d_x(eP^*_X)] - P^*_M C_m(p_h, eP^*_M, z) = B(p_h, e, m, k)$$

E. *The Labor Market*

The excess demand for labor L can be written as

$$(9) \quad L = l^d_x(eP^*_X) + l^d_h(p_h) - \overline{L} = L(p_h, e, k)$$

Let us now look at the budget constraint. Rewrite equation (5) as

$$(10) \qquad y = p_h C_h + eP^*_M C_m + \Pi$$

With the help of equation (3), we can rearrange the above equation into

$$(11) \qquad eB = p_h H + \Pi$$

This means that if any two of the three effective markets (B, H, Π) clear, the last effective market must automatically clear, even though the labor market may not.

The model deals with four markets (Π, B, H, L), three real variables (p_h, e, m), and one exogenous variable k. At short-run equilibrium (or at the end of Stage I adjustments), the three effective markets (Π, B, H) clear though the labor market does not. As for the fixed or free exchange-rate regime, the two endogenous variables are either (p_h, m) or (p_h, e), respectively; since, the respective variable $e = \rho/W$ or $m = M^s/W$ is considered as fixed at Stage I. The budget equation (11) tells us that we can drop one market, leaving us with a determinate system of two endogenous variables and two market-clearing equations. At long-run equilibrium (or at the end of Stage II adjustments), all four markets clear. Again, from the budget constraint (11) we can drop the market-equilibrium equation when we solve for the three endogenous variables (p_h, e, m). In a system of free or fixed exchange rates, the variables in the long run

FIGURE 1. THE FREE EXCHANGE-RATE REGIME

FIGURE 2. THE FIXED EXCHANGE-RATE REGIME (For the sake of illustration, the functions are drawn with the assumption that imports and home goods are gross substitutes for each other.)

are either (ρ, P_H, W) or (M^s, P_H, W), respectively. However, the homogeneity property of the model allows us to rewrite the excess demand equations in terms of relative price and real balances $(M^s/W, \rho/W, P_H/W)$ for both exchange-rate regimes. Therefore, at the long-run equilibrium, both fixed and free exchange-rate regimes must have identical values for all *real* variables—output, employment, relative prices, and real money balances.

F. *A Diagrammatic Exposition of the Model*

Let us illustrate the model by means of Figures 1 and 2. Assume an exogenous disturbance (say, a change in k) shifts the final equilibrium from 0 to F in Figures 1 and 2. For the free or fixed exchange-rate regime we employ either a (p_h, e) or (p_h, m) space, respectively, for our analysis. The $H = 0$, $\Pi = 0$, and $L = 0$ are the market-clearing loci of equations (4′) and (7)–(9). During Stage I adjustment, W is fixed; either (p_h, e) or (p_h, m)[11] adjust rapidly in the respective regimes to clear the home good market and the market of foreign exchange. At point T in Figures 1 and 2, Stage I adjustment concludes. The gradual adjustment of the money wage in Stage II acts as a shifting parameter following the completion of Stage I adjustment. The movement of W can be traced by observing that $m = M^s/W$ or $e = \rho/W$ can be a proxy for W in a free or fixed exchange-rate regime since M^s or ρ is constant. In other words, we can portray short- and long-run dynamic adjustments in (p_h, e) or (p_h, m) space with m or e as a shift parameter in the respective background; by varying m or e, we shift the market-clearing loci towards the final equilibrium point F. The H^L lines[12] in Figures 1 and 2 are the locus of equilibrium points traced out from the intersections of $H = 0$, $\Pi = 0$, and $B = 0$ by varying the respective shift parameter m or e. Hence the H^l line can be referred to as the time path at Stage II, or the long-run time path of adjustment.

II. The Short-Run Level of Unemployment

At short-run equilibrium (or at the end of Stage I adjustment), the level of unemployment u is equal to the excess supply of labor $-L$. We therefore introduce one more variable u and one additional equation into the system; i.e.,

$$(12)\qquad \begin{aligned} H &= 0 \\ B &= 0 \\ L + u &= 0 \end{aligned}$$

For the fixed or free exchange-rate regime, we can solve for the three endogenous variables either (u^l, p_h^l, m^l) or (u^r, p_h^r, e^r), respec-

[11]Note that with W fixed at Stage I, (p_h, e) or (p_h, m) can be proxy for either (P_H, ρ) or (P_H, M^s), respectively

[12]The slope of the H^L line is always positive (see the Appendix.).

tively, in terms of the exogenous variable k. In the following lemma, an analytical relationship between short-run unemployment and long-run net change in the wage rate will be established.

LEMMA 1: *The net change in the level of unemployment at short-run equilibrium due to a change in the exogenous parameter k is (negatively) related to the net change in the level of money wage at long-run equilibrium in the following ways:*

$$u^i_k = \left(\frac{e\Delta}{W\Delta^i}\right) W^i_k \tag{13}$$

$$u^r_k = \left(\frac{-m\Delta}{W\Delta^r}\right) W^r_k \tag{14}$$

where $u^i_k = du^i/dk$, $u^r_k = du^r/dk$, $W^i_k = dW^i/dk$, $W^r_k = dW^r/dk$ *and the* Δ *terms denote determinants.* (For notation and proof see the Appendix.)

III. The Long-Run Time Paths

In this section, we examine the shapes of the long-run time paths of variables as they adjust toward their long-run equilibrium values. In order to have a workable dynamic comparison of the two regimes, assume that the foreign exchange market B and the home goods market H adjust *instantaneously* while money wage adjusts slowly in proportion to the excess demand in the labor market L.[13] In other words, for an exogenous disturbance, the overall time path of each variable is characterized by an initial jump at t = 0 due to the above assumption of instantaneous Stage I adjustments. When t > 0, each variable adjusts gradually and must satisfy the equation system (15) below under both regimes:

$$(15) \qquad \begin{aligned} DW &= \theta L \\ 0 &= H \\ 0 &= B \end{aligned}$$

The free exchange-rate regime adjusts in the following manner:

$$(16) \qquad \begin{aligned} DW &= \theta L[P_H(t), \rho(t), W(t), k] \\ 0 &= H[P_H(t), \rho(t), W(t), k] \\ 0 &= B[P_H(t), \rho(t), W(t), k] \end{aligned}$$

The characteristic root $(-\lambda^r)$ of (16) can be obtained by solving the following characteristic equations (see the Appendix):

$$(17) \qquad \Delta(\lambda^r) = \begin{vmatrix} H_{P_H} & H_\rho & H_W \\ B_{P_H} & B_\rho & B_W \\ \theta L_{P_H} & \theta L_\rho & \theta L_W - (-\lambda r) \end{vmatrix} = 0$$

$$-\lambda^r = -\frac{m}{W}\frac{\theta\Delta}{\Delta^r} < 0 \tag{18}$$

Since the short-run level of unemployment is equal to $u^r_k \Delta k$, and the long-run equilibrium level of unemployment is zero, the long-run time path of unemployment $u^r(t)$ is

$$u^r(t) = (u^r_k \Delta k) e^{-\lambda^r t} \tag{19}$$

Similarly, the fixed exchange-rate regime adjusts in the following manner:

$$(20) \qquad \begin{aligned} DW &= \theta L[P_H(t), W(t), k] \\ 0 &= H[P_H(t), M^s(t), W(t), k] \\ 0 &= B[P_H(t), M^s(t), W(t), k] \end{aligned}$$

The characteristic root λ^i can be obtained by solving the characteristic equations of (20). We then obtain

$$-\lambda^i = \frac{\Delta}{\Delta^i}\frac{\theta e}{W} < 0 \tag{21}$$

Similarly, the time path of unemployment $u^i(t)$ can be written

$$u^i(t) = (u^i_k \Delta k) e^{-\lambda^i t} \tag{22}$$

Since there is only one characteristic root associated with the dynamic systems of the

[13]This assumption can be justified as follows: In the conventional wisdom, economists (for example, Friedman) generally agree that if internal wages were as flexible as exchange rates, it would make very little economic difference whether adjustments were brought about by a free exchange-rate regime or a fixed exchange-rate regime. In order to highlight the effects of a sticky money wage, I disregard the difference in adjustment speeds between fixed and free exchange-rate regimes in the market for foreign exchange, and between the home goods market and the market of foreign exchange. I simply assume that P_H and ρ (or M^s) adjust to clear their respective markets instantaneously.

fixed and free exchange-rate regimes, there can be no overshooting during either adjustment process. Each system's speed of response is determined by its particular λ, which in turn is determined by the magnitude of the coefficients in the excess demand equations.

Let us now examine the total loss of employment over an infinite time horizon for both regimes.

THEOREM 1: *The total loss in employment over an infinite time horizon caused by an exogenous disturbance* (k) *is proportional to the net change of money wages for that particular regime, i.e.,*

$$\int_0^\infty u^r(t)\,dt = \frac{1}{\theta} W^r_k \Delta k \tag{23}$$

$$\int_0^\infty u^i(t)\,dt = \frac{1}{\theta} W^i_k \Delta k \tag{24}$$

PROOF:

Equation (23) of the free exchange-rate regime is equivalent to the following:

$$\int_0^\infty u^r(t)\,dt = \int_0^\infty u^r_k \Delta k\, e^{-\lambda^r t}\,dt = \frac{u^r_k \Delta k}{\lambda^r} \tag{25}$$

Substituting equations (18) and (14) of Lemma 1 into (25), we get

$$\frac{u^r_k}{\lambda^r} \Delta k = \frac{1}{\theta} W^r_k \Delta k \tag{25$'$}$$

And, substituting the above equation into (25) we get equation (23). Equation (24) of the fixed exchange-rate regime can also be proven in the same way.

The results obtained in Theorem 1 are due to assumptions made in the equation system (15), which applies to both regimes. Since the home goods market and foreign exchange market are assumed to clear instantaneously, we can therefore solve for (ρ, P_H) or (M^s, P_H) in terms of W from the two equations $H = 0$ and $B = 0$. Substituting ($\rho(W)$, $P_H(W)$) or ($(M^s(W), P_H(W)$) into the relevant excess demand for labor equation L, we get a differential equation of one variable W alone, i.e., $DW = \theta L(W)$. Thus, integrating both sides of this differential equation, we get the results in Theorem 1. Let us now proceed to the dynamic evaluation of exchange-rate regimes in the following section.

IV. Dynamic Evaluation of Exchange-Rate Regimes

The Stein criterion can be reinterpreted in terms of the familiar transformation equation of variables between regimes. Kemp first noted this "family" relationship of systems. The long-run equilibrium wage rate W is related between fixed and free regimes in the following way:

$$W^{rf}_k = W^{if}_k - W^{if}_\rho \cdot (M^{if}_\rho)^{-1} \cdot M^{if}_k \tag{26}$$

where the superscript (f) indicates the variables considered are evaluated at long-run equilibrium. Hence, W^{rf}_k is the long-run equilibrium change in the money wage from an exogenous disturbance under free exchange rates. The signs of W^{if}_ρ and M^{if}_ρ are positive (see the Appendix).

A conflict (compatible) economy can be interpreted in our framework as one where the sign of the term ($W^{if}_k \cdot M^{if}_k$) is greater (less) than zero.[14] When *sgn* (W^{if}_k) = *sgn* (W^{rf}_k), we can conclude from equation (26) that a free (fixed) exchange-rate regime is superior for a conflict (compatible) economy.[15]

We can further extend and strengthen Stein's criterion to a dynamic setting by the following theorem:

THEOREM 2: *For a conflict* (*compatible*)

[14]Stein's concept of conflict/compatible economy is based on a short-run Keynesian framework of analysis. He defines a conflict (compatible) economy as one where the balance of payments tends towards surplus (deficit) during the expansion (contraction) phase of the business cycle. Stein's definition of a conflict (compatible) economy can be reinterpreted in our disequilibrium framework, as one where the sign of ($u^i_k \cdot M^i_k$) is less (greater) than zero after the completion of the instantaneous Stage 1 adjustments and/or the sign of ($W^{if}_k \cdot M^{if}_k$) is greater (less) than zero after the completion of long-run adjustments (see Lemma 1).

[15]Note that Stein's criterion is by no means general because it requires *sgn* (W^{if}_k) = *sgn* (W^{rf}_k). This means the depreciation of currencies will not be strong enough to raise the level of employment to the region of excess demand for labor. Stein dismisses this possibility as "most unlikely to occur."

FIGURE 3 CONFLICT ECONOMY

economy, the net loss in employment of a fixed (free) exchange-rate regime over an infinite time horizon is always greater than that of a free (fixed) exchange-rate regime even if the former regime has a faster speed of adjustment than the latter.

PROOF:

The proof to Theorem 2 follows from Theorem 1. The above intrepretation of a conflict economy implies $|W_k^{i\prime}| > |W_k^{r\prime}|$, and for a compatible economy $|W_k^{i\prime}| < |W_k^{r\prime}|$; Theorem 1 tells us that the net loss in employment over an infinite time horizon is proportional to the net change in money wage; therefore, the net loss in employment of a fixed (free) exchange-rate regime over an infinite time horizon must be greater than the free (fixed) exchange-rate regime. (Theorem 2 is illustrated in Figures 3 and 4.)

In Figure 3 the case of a conflict economy is shown. The exogenous disturbance k comes at $t = 0$. There is an immediate jump in the level of unemployment at $t = 0$, due to our assumption of instantaneous Stage I adjustments. Then unemployment gradually declines due the gradual adjustment of money wage. The implication of Theorem 2 is that even if $\lambda^i > \lambda^r$, the total area under the curve $u^i(t)$ is always greater than the total area under the curve $u^r(t)$. This means the shaded area A in Figure 3A is always greater than the shaded area B. The case where $\lambda^i < \lambda^r$ is

FIGURE 4. COMPATIBLE ECONOMY

shown in Figure 3B. Figures 4A and B show the case of a compatible economy.

A. *The Welfare Implications of Unemployment*

As the level of unemployment gradually declines with time, the production possibility frontier (S_x, S_h) gradually expands. Since export and import prices are fixed in the rest of the world, we can multiply the S_x axis by P_X^*/P_M^* and transform it into the consumption possibility frontier (S_m, S_h). Our assumption of instantaneous adjustment of product prices implies that the marginal rate of substitution must always equal the marginal rate of transformation. Therefore, it is unambiguous in our model that a lower level of unemployment must have a higher level of welfare since the consumption possibility frontier is now bigger.

Suppose we now construct a social cardinal utility function with diminishing marginal utility on the level of employment and a time rate of discount. As for the conflict economy, it is obvious that when $\lambda' > \lambda'$ (as shown in Figure 3B), our main conclusion still holds; as for the case where $\lambda' > \lambda'$ (Figure 3B), given a discount rate, future losses in unemployment do not carry as much weight. Hence, the shaded area B in Figure 3A should be weighted smaller. Diminishing marginal utility also means that the shaded area B should again be weighted less since it is closer to full employment than the shaded area A. Thus, for the case of a conflict economy, our previous conclusion holds. We can construct the same kind of argument for the case of a compatible economy.

V. Conclusion and Extension of the Model

My construction of a disequilibrium model and the assumption of a two-stage adjustment process enables us to compare the time paths of involuntary unemployment of different exchange-rate regimes. The analysis shows that Stein's criterion can be extended to a dynamic setting. It can also be strengthened as follows: for a conflict (compatible) economy, a free (fixed) exchange-rate regime is superior to a fixed (free) exchange-rate regime even though the former regime has a slower speed of adjustment than the latter. However, as mentioned in footnote 14, Stein's criterion does not include all the possible cases. To analyze those cases, one may have to know the specific magnitude of coefficients and the specific form of the welfare function before any comparisons can be made.

Our model assumes implicitly that the money wage is sticky in a downward direction and flexible like any other price in an upward direction. Hence, in the region of excess demand for labor, it makes very little economic difference whether adjustments are brought about by fixed or free exchange-rate regimes. The only relevant part is when exogenous disturbances throw the economy into the region of excess supply of labor. However, if we are willing to assume flexible commodity prices but an inflexible money wage during periods of labor shortage, the case of excess demand for labor becomes important. Fortunately, our model and results can easily be extended to the case of excess demand for labor without loss of generality (see the Appendix).

APPENDIX

Assume that four excess demand equations (H, B, Π, L) shown earlier can be linearized. Let

$$\Delta' = \begin{vmatrix} H_{p_h} & H_e \\ B_{p_h} & B_e \end{vmatrix} \quad \Delta' = \begin{vmatrix} H_{p_h} & H_m \\ B_{p_h} & B_m \end{vmatrix}$$

$$\Delta_1 = \begin{vmatrix} H_e & H_m \\ B_e & B_m \end{vmatrix} \quad \Delta = \begin{vmatrix} H_{p_h} & H_e & H_m \\ B_{p_h} & B_e & B_m \\ L_{p_h} & L_e & 0 \end{vmatrix}$$

To calculate the values of the above determinants, we have to make use of the following: 1) the budget constraint $eB = p_hH + \Pi$, so that we can substitute the partial derivatives of B (with respect to (p_h, e, m, k)) by partial derivatives of H and Π; 2) we have to assume the economy is near the region of equilibrium so that $H \simeq 0$ $B \simeq 0$; and 3) we have to employ Slustsky's decomposition of

demand functions into substitution and income terms, to derive the following:

$$\text{(A1)}\quad \Delta^r = \Pi_y\Big(\underset{(-)}{r_1}\,\underset{(+)}{\frac{dy}{de}}\,\frac{1}{P_M} - \underset{(+)}{r_2}\,\underset{(+)}{\frac{dy}{dp_h}}\Big) < 0$$

$$\text{(A2)}\quad \Delta^l = \underset{(-)}{\Pi_m}\underset{(+)}{r_1} > 0$$

$$\text{(A3)}\quad \Delta^1 = \underset{(-)}{\Pi_m}\underset{(+)}{r_2} < 0$$

$$\text{(A4)}\quad \Delta = \underset{(\ )}{\Pi_m}\Big[-e\frac{P_X^*}{P_M^*}\underset{(+)}{\frac{dS_x}{de}}\Big(\underset{(\ )}{\frac{dC_h}{dp_h}\Big|_s} - H\underset{(+)}{\frac{dC_h}{dz}} - \underset{(+)}{\frac{dS_h}{dp_h}}\Big) + \frac{p_h}{P_M^*}\underset{(+)}{\frac{dS_h}{dp_h}}\Big(\underset{(+)}{\frac{dC_h}{de}\Big|_s} + B\underset{(+)}{\frac{dC_h}{dz}}\Big)\Big] < 0$$

where

$$\text{(A5)}\quad r_1 = \underset{(\ )}{\frac{dC_h}{dp_h}\Big|_s} - H\underset{(+)}{\frac{dC_h}{dz}} - \underset{(+)}{\Big(1 - p_h\frac{dC_h}{dz}\Big)}\frac{dS_h}{dp_h} < 0$$

$$\text{(A6)}\quad r_2 = \frac{1}{P_M^*}\Big(\underset{(+)}{\frac{dC_h}{de}\Big|_s} + B\underset{(+)}{\frac{dC_h}{dz}}\,eP_X^*\underset{(+)}{\frac{dS_x}{de}}\Big) > 0$$

The signs in parentheses are the signs of the corresponding derivatives. The subscript s indicates the term is the substitution term in the Slustsky equation. Hence, $dC_h/dp_h|_s < 0$ and $dC_m/de|_s < 0$. Our assumption that consumers are free of money illusion (see fn. 9) implies $d\Pi/de|_s = 0$. And, from consumer theory, we know that $p_h(dC_h/de)|_s + e(dC_m/de)|_s + d\Pi/de|_s = 0$ (see James Henderson and Richard Quandt, p. 39); therefore $dC_h/de|_s$ must be positive.

In Figures 1 and 2, the H^L line is positively sloped. For the fixed exchange-rate regime, we can solve $H = 0$ and $\Pi = 0$ for dp_h/de, and dm/de. The slope of the H^L line in the (p_h, m) space is therefore equal to $(dp_h/de)/(dm/de)$, i.e., $dp_h/dm = \Delta_1/\Delta^r > 0$.

For the free exchange-rate regime, the slope of the H^L line in the (p_h, e) space is equal to $(dp_h/dm)/(de/dm) = -\Delta_1/\Delta^l > 0$

PROOF of Lemma 1:

To prove Lemma 1, let us rewrite equation (12):

$$\text{(12')}\quad \begin{aligned} H(p_h, e, m, k) &= 0 \\ B(p_h, e, m, k) &= 0 \\ L(p_h, e, k) + u &= 0 \end{aligned}$$

For the fixed exchange-rate regime we can solve u_k^l (the derivative of u^l with respect of k) of equation (12') above by Cramer's rule:

$$\text{(A7)}\quad u_k^l = \frac{1}{\Delta^l}\begin{vmatrix} H_{p_h} & H_m & -H_k \\ B_{p_h} & B_m & -B_k \\ L_{p_h} & 0 & -L_k \end{vmatrix}$$

For a free exchange-rate regime, we get:

$$\text{(A8)}\quad u_k^r = \frac{1}{\Delta^r}\begin{vmatrix} H_{p_h} & H_e & -H_k \\ B_{p_h} & B_e & -B_k \\ L_{p_h} & L_e & -L_k \end{vmatrix}$$

At long-run equilibrium, we can solve $H = 0$, $B = 0$, $L = 0$ for (p_h, e, m). Note that both regimes must have identical real variables, hence, identical values of (p_h, e, m). Solving by Cramer's rule, we get:

$$\text{(A9)}\quad e_k = \frac{1}{\Delta}\begin{vmatrix} H_{p_h} & -H_k & H_m \\ B_{p_h} & -B_k & B_m \\ L_{p_h} & -L_k & 0 \end{vmatrix}$$

and

$$m_k = \frac{1}{\Delta}\begin{vmatrix} H_{p_h} & H_e & -H_k \\ B_{p_h} & B_e & -B_k \\ L_{p_h} & L_e & -L_k \end{vmatrix}$$

For the fixed (free) exchange-rate regime, since $\rho(M^s)$ is fixed, $e(m)$ can be a proxy for W. Hence,

$$\text{(A10)}\quad m_k = -\frac{m}{W}W_k'$$

$$\text{(A11)}\quad e_k = -\frac{e}{W}W_k'$$

By substituting equations (A10) and (A11) into (A7)–(A9), we get the result of Lemma 1.

To solve equations (17) and (20) for λ^r and λ^i we need to note the following relationships:

$$\text{(A12)} \quad H_{P_H} = H_{p_h}\frac{1}{W}$$

$$\text{(A13)} \quad H_{M'} = H_m\frac{1}{W}$$

$$\text{(A14)} \quad H_P = H_e\frac{1}{W}$$

$$\text{(A15)} \quad H_W = -\frac{1}{W}(H_{p_h}p_h + H_e e + H_m m)$$

As for equation (26), we can derive the long-run effects of devaluation on (W^{ff}, M^{ff}) of the fixed exchange-rate regime. By using Cramer's rule to solve the three market-clearing equations ($H = 0$, $B = 0$ and $L = 0$) and by applying the relations (A12) to (A15) (similar relationships can be derived for B and L), we can get:

$$\text{(A16)} \quad \frac{dW}{d\rho} = W^{ff}_\rho = \frac{\Delta}{W^3}\Big/\frac{e\Delta}{W^3} = \frac{1}{e} > 0$$

$$\text{(A17)} \quad \frac{dM'}{d\rho} = M^{ff}_\rho = \frac{m\Delta}{W^3}\Big/\frac{e\Delta}{W^3} = \frac{m}{e} > 0$$

The Excess Demand for Labor Case

We can again assume a two-stage adjustment process as before. Voluntary exchange under disequilibrium implies that the representative firm perceives a supply-imposed constraint on employment. Assume that the effective labor supplies, l^s_x and l^s_h, of the export and home goods markets obey the following rudimentary rules:

$$\text{(A18)} \quad l^s_x = \beta l^d_x(P^*_X \cdot e) + \alpha[\bar{L} - l^d_h(p_h)] = l^s_x(p_h, e, k)$$

$$\text{(A19)} \quad l^s_h = \alpha l^d_h(p_h) + \beta[\bar{L} - l^d_x(P^*_X \cdot e)] = l^s_h(p_h, e, k)$$

where $\alpha + \beta = 1$, $\alpha > 0$, $\beta > 0$ and α and β are arbitrary constants. I argue that this is the most general allocation rule. Due to the labor supply constraint, the equations for effective real income, expenditure, and accumulation of money balances take the following forms:

$$\text{(A20)} \quad y = p_h S_h(l^s_h) + eP^*_X S_x(l^s_x) = y(p_h, e, k)$$

$$\text{(A21)} \quad \Pi = \Pi(y, m)$$

$$\text{(A22)} \quad z = y - \Pi = z(p_h, e, m, k)$$

On the demand side of home goods market, we have to employ effective income y as formulated in equation (A20). On the supply side, the supply of home goods is constrained by the effective labor supply l^s_h.

$$\text{(A23)} \quad H = C_h(p_h, eP^*_M, z) - S_h(l^s_h) = H(p_h, e, m, k)$$

Similarly, import demand is formulated in terms of effective income. The supply of exports is based upon effective labor supply l^s_x. Hence, the excess demand function for foreign exchange is

$$\text{(A24)} \quad B = P^*_X S_x(l^s_x) - P^*_M C_m(p_h, eP^*_M, z) = B(p_h, e, m, k)$$

Since all the excess functions (A21), (A23), and (A24) are identical in form to the case of excess supply of labor, Lemma 1 and Theorems 1 and 2 must also hold for the case of excess demand for labor. Hence, we can draw the same conclusions here as in the case of excess supply of labor.

REFERENCES

Robert J. Barro and Herschel I. Grossman, *Money, Employment and Inflation*, Cambridge 1976.

______ **and** ______, "A General Disequilibrium Model of Income and Employment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 82–93.

D. L. Brito and J. D. Richardson, "Some Disequilibrium Dynamics of Exchange Rate Changes," *J. Int. Econ.*, Jan. 1975, *5*, 1–13.

K. S. Chan, "Exchange Rates and Tariffs in a Non-Market-Clearing Model of International Finance," unpublished doctoral dissertation, Brown Univ. 1977.

R. W. Clower, "The Keynesian Counterrevolution: A Theoretical Appraisal," in Frank Hahn and Frank Brechling, eds., *The Theory of Interest Rates*, London 1965.

M. Connolly and D. Taylor, "Adjustment to Devaluation with Money and Nontraded Goods," *J. Int. Econ.*, Aug. 1976, *6*, 289–98.

R. Dornbusch, "Expectations and Exchange-Rate Dynamics," *J. Polit. Econ.*, Dec. 1976, *84*, 1161–76.

M. Friedman, "The Case for Flexible Exchange Rates," in his *Essays in Positive Economics*, Chicago 1953, 157–203.

H. I. Grossman, "Money, Interest and Prices in Market Disequilibrium," *J. Polit. Econ.*, Sept./Oct. 1971, *79*, 943–61.

James M. Henderson and Richard E. Quandt, *Microeconomic Theory: A Mathematical Approach*, 2d ed., New York 1971.

Murray C. Kemp, *The Pure Theory of International Trade*, Englewood Cliffs 1964.

S. Laursen and L. A. Metzler, "Flexible Exchange Rates and the Theory of Employment," *Rev. Econ. Statist.*, Sept. 1950, *32*, 281–99.

Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, New York, 1968.

R. F. Lucas, "Monetary Union: An Alternative to Free Exchange Rates?," unpublished doctoral dissertation, Brown Univ. 1975.

Edmond Malinvaud, *The Theory of Unemployment Reconsidered*, New York 1977.

Robert A. Mundell, *International Economics*, New York 1968.

Don Patinkin, *Money, Interest and Prices: An Integration of Monetary and Value Theory*, 2d ed., New York 1965.

Egon Sohmen, *Flexible Exchange Rates: Theory and Controversy*, Chicago 1961.

J. L. Stein, "The Optimum Foreign Exchange Market," *Amer. Econ. Rev.*, June 1963, *53*, 384–402.

Black-White Life Cycle Earnings Differences and the Vintage Hypothesis: A Longitudinal Analysis

By SAUL D. HOFFMAN*

A common finding of virtually all previous studies of black-white earnings differences is that the earnings of white males appear to rise at a more rapid rate with each additional year of work experience (or with age) than do the earnings of black workers. This finding has been noted by economists since at least the mid-1950's when Morton Zeman called attention to the declining pattern of black-white income ratios by age cohort in the 1940 Census. It has been confirmed both with the more recent census data and by the many recent studies of black and white earnings based on survey data. The obvious implication of this finding is that black-white differences tend to increase over the life cycle of an individual and, moreover, that the labor market itself may be a vehicle for generating these earnings differences.

This issue of black-white life cycle earnings differences is closely tied to economic models of labor market earnings, since most of these models place primary emphasis on the explanation of the life cycle pattern of individual earnings.[1] There are at least two general explanations of increasing black-white earnings differences over the life cycle, each corresponding to a different model of the labor market. Either discrimination itself may be constant over the life cycle, but the relative productivities of black and white workers—and hence, their earnings—may diverge over the life cycle or, conversely, relative individual productivities may be constant, but discrimination may increase over the life cycle. The first explanation is along the lines of the human capital model, while the second is best captured by dual or segmented labor market models.

The basic empirical finding of increasing life cycle earnings differences is, however, questionable, since it has been drawn almost exclusively from cross-sectional rather than longitudinal data. The apparent life cycle result has actually been inferred from the earnings of different individuals of different ages at a single point in time, rather than from the earnings of a single individual over the life cycle. It is possible, then, that the cross-sectional result might misrepresent the life cycle earnings profile of black and white workers and that the finding of increasing life cycle earnings differences might be simply a statistical artifact of the cross section. Welch (1973a, b) has made this argument in a series of recent papers, citing what he calls "vintage effects"—the relative differences between older black and white workers and between younger ones—as a plausible explanation of the cross-sectional finding.[2] Formally, Welch's vintage hypothesis implies that the cross-sectional model is actually a misspecified model of life cycle earnings.

This paper reexamines the issue of black-white life cycle earnings, paying attention both to alternative theories of life cycle earnings and to the implications of the vintage hypothesis for empirical work. In order to test the vintage hypothesis and to estimate black-white life cycle earnings patterns, a cross-

*Assistant professor of economics, University of Delaware. I would like to acknowledge the many suggestions I received from the members of my thesis committee, George Johnson, Frank Stafford, and William Birdsall, and from the Staff of the Panel Study of Income Dynamics, especially James Morgan and Greg Duncan. The presentation of this paper was greatly improved by the comments of the referees and the managing editor. Financial assistance from a University of Delaware General University Research Grant is gratefully acknowledged.

[1]For example, Finis Welch concluded in his summary of the human capital model that "life-cycle earnings patterns are the core of this theory" (1975, p. 72).

[2]James Smith and Welch provide an updated empirical analysis of the vintage hypothesis.

sectional earnings function and a pooled cross-section and time-series earnings function are estimated using longitudinal data on individual earnings from the *Panel Study of Income Dynamics*. The pooled results provide an improved estimate of life cycle earnings differences while the comparison of the cross-sectional and pooled results provides a test of the vintage hypothesis.

Section I briefly reviews several models of life cycle earnings and summarizes their implications for black-white life cycle earnings differences. Section II turns to the previous empirical work in this area and then outlines the vintage hypothesis and its econometric implications for the cross-sectional estimation of life cycle effects. Section III discusses the problems encountered in the estimation of a life cycle earnings model. In the absence of a sufficiently large sample of individuals, a pooled cross-section and time-series earnings model is developed in which individuals from several vintages, defined by age, are pooled together for estimation. Finally, Section IV presents cross-sectional earnings functions and pooled earnings functions for two cohorts of black and white working males. Section V provides a brief summary of the paper and considers its policy implications.

I. Life Cycle Earnings Models

According to the human capital model (see Gary Becker; Yoram Ben-Porath; Sherwin Rosen), the life cycle pattern of individual labor market earnings reflects that individual's life cycle pattern of productivity which, in turn, depends on the time profile of self-investments in human capital. At every point in time, the decision to invest in training depends on the costs and benefits of acquiring an additional unit of training, where the benefits are a function of the rental rate of human capital and the length of the payoff period over which the returns to the investment can be accrued. The costs are measured by the potential earnings which are sacrificed in order to acquire training. The familiar result of these models is that investments should be concentrated in an individual's early working years in order to maximize the number of years over which earnings can be accrued, and that thereafter the volume of investments should taper off. The empirical counterpart of this life cycle investment profile is a concave experience-earnings profile, since earnings are assumed to be depressed initially by the large volume of investment, but rise subsequently as investments decline and the returns to previous investments are received.

In a life cycle context, the most important form of investment in human capital is usually assumed to be investment in on-the-job training. The important question with respect to black-white life cycle earnings differences is whether the relevant parameters of the human capital investment function differ significantly by race, so that blacks might have economic incentives to acquire less training. There are at least two possibilities. One is labor market discrimination which lowers the value of any investment and might reduce the incentive to acquire job skills. However, this result is ambiguous, since in some models discrimination reduces the costs of investment (measured by the opportunity cost of foregone earnings) by an equivalent amount, leaving the optimal volume of investment unaffected.[3] The other possible explanation draws on differences by race in the quality of education. In most human capital models, the costs of investment are assumed to be a decreasing function of an individual's ability to learn. Since blacks tend to have less education than whites and, perhaps, poorer quality education as well, their investment costs might be greater and, hence, they would have an incentive to acquire less human capital. Thus, it is conceivable that black-white skill differences

[3]Discrimination has no effect in a Ben-Porath-type model in which an individual's own stock of human capital is the only input in the human capital production function. It reduces optimal investment in Rosen's model, since the cost of training is a market-determined price rather than the opportunity cost of time spent acquiring training. The effects of discrimination on optimal investment in the utility-maximization models (see Harl Ryder, Frank Stafford, and Paula Stephan; Alan Blinder and Yoram Weiss; James Heckman) are ambiguous, since income and substitution effects may conflict.

might increase over the life cycle as a result of differential rates of investment in on-the-job training.

In contrast, in the dual labor market model, life cycle earnings are thought to be largely determined by the labor market in which an individual works rather than the skills (or human capital) he possesses. The labor market is assumed to consist of two distinct markets, a primary sector and a secondary sector. For all the much-criticized vagueness of the model concerning the distinguishing features of jobs in the two sectors, it is probably reasonable to think of primary sector jobs as "jobs with a future" and secondary sector jobs as "dead-end jobs." Mobility between the sectors is limited, and it is alleged that one of the primary manifestations of labor market discrimination is to relegate blacks to secondary labor market jobs in disproportionate numbers relative to their skills. Moreover, those blacks who can find primary sector employment are assumed to face discrimination in promotions.[4] The result is that earnings differentials by race would be expected to increase over the life cycle.

It is not possible in empirical work to distinguish between the two human capital model explanations (discrimination and educational differences) or between the human capital model and the dual labor market explanation. The problem is that only the outcome (an experience-earnings profile) rather than the postulated underlying processes (investment in on-the-job training or the mechanisms of occupational segregation) are directly observable.

II. Cross-Sectional Studies and the Vintage Hypothesis

Since the mid-1960's, most analyses of black-white earnings differences have been based on cross-sectional survey data which provided information on individual earnings as well as basic demographic data and, depending on the data set, other individual characteristics as well. In general, these studies (see, for example, Giora Hanoch; Robert Hall; Blinder) estimate separate earnings functions for blacks and whites and then examine differences in the estimated coefficients. Virtually all of these studies find clear evidence of increasing black-white earnings differences with respect to age or years of labor market experience in the form of lower regression coefficients for blacks than for whites. This is true regardless of the data source used or the set of independent variables included in the analysis. Indeed, the evidence is so strong that Hall concluded from his analysis of the "Survey of Economic Opportunity" (*SEO*) data that "the whole notion of a career with steady advancement is relevant only for white males" (pp. 393–94).

One problem common to all of these studies is that they have attempted to infer individual life cycle earnings patterns by constructing a "synthetic life cycle" from cross-sectional data on the earnings of different individuals of different ages. A procedure of this kind is appropriate only if the factors which affect the relative earnings of blacks have been constant over the time period covered by the cross section. If there are time-related factors which have affected the relative earnings of blacks and whites but are not explicitly incorporated into the cross-sectional equation, then the cross-sectional equation will be misspecified and the estimated coefficients will be biased. This bias will be especially serious for a variable like years of work experience, itself time-related and therefore strongly correlated with the omitted time-related factors.

An argument of this kind is implicit in Welch's vintage hypothesis. He argues that the cross-sectional results of previous studies reflect not the life-cycle effects of discrimination, but rather vintage effects in the form of an improvement over time in the relative quality of schooling available to blacks and/or a secular decline in discrimination. Thus, the lower relative earnings of older black workers might result either from large differences between older blacks and older whites in the quality of education or from the lingering effects of the more severe labor market discrimination which older blacks

[4]An argument of this kind is also made by Kenneth Arrow and by Rosen

faced when the first entered the labor market.

In his own work, Welch has emphasized improvements over time in the quality of education available to black school children, particularly in the South. He cites, for example, a number of nominal input measures of educational quality, such as per capita expenditures, pupil-teacher ratios, and length of school year, all of which have increased relatively more rapidly for blacks than whites during the twentieth century. Thus, he argues that younger black workers—who are of more recent vintages and who therefore received a better quality education—are of higher quality relative to whites than are older blacks, and consequently they receive relatively higher wages.

It is useful to formalize the vintage hypothesis in order to highlight its implications for the cross-sectional estimation of black-white life cycle earnings differences. According to Welch's vintage hypothesis, earnings are a function not only of the conventional cross-sectional explanatory variables, but of two other variables as well: the quality of an individual's education and the extent and strength of discrimination at the time that individual entered the labor market. We can express this model as

$$(1)\quad Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \beta_Q Q_i + \beta_D D_i + \epsilon_i$$

where the X_{ij} are conventional explanatory variables, Q_i is a measure of the quality of an individual's education, and D_i is a summary measure of the prevailing labor market discrimination at the time an individual entered the labor market; $\beta_Q > 0$ and $\beta_D < 0$. Both Q and D are in turn hypothesized to be functions of the time period in which an individual attended school and first entered the labor market; in Welch's terms, they are functions of an individual's vintage:

$$(2)\quad Q_i = \alpha_1 V_i + \eta_i$$

$$(3)\quad D_i = \alpha_2 V_i + \mu_i$$

Presumably $\alpha_1 < 0$ (school quality was poorer for older vintages) and, for blacks, $\alpha_2 > 0$ (labor market discrimination was more severe for older vintages).

In most cross-sectional analyses, however, Q and D are omitted, and the estimated equation is

$$(4)\quad Y_i = \beta_0 + \sum_{j=1}^{K} \beta_j X_{ij} + \epsilon_i^*$$

where $\epsilon_i^* = \epsilon_i + \beta_Q Q_i + \beta_D D_i$. The cross-sectional equation is therefore misspecified.

Rewriting (4) in matrix notation as $Y = X^*\beta^* + \epsilon^*$ and solving for $\hat{\beta}^*$, we have

$$(5)\quad \hat{\beta}^* = (X^{*\prime}X^*)^{-1}X^{*\prime}Y$$

Substituting for Y from (1) and for Q and D from (2) and (3), and then taking expectations, we can express the vector of estimated cross-sectional coefficients as

$$(6)\quad E(\hat{\beta}^*) = \beta^* + (\beta_Q\alpha_1 + \beta_D\alpha_2) \cdot E[(X^{*\prime}X^*)^{-1}X^{*\prime}V]$$

The bracketed expression in (6) can be thought of as the (K x 1) vector of "coefficients" from the "regression" of vintage on the entire set of included variables:[5]

$$(7)\quad V_i = \gamma_0 + \gamma_1 X_{i1} + \ldots + \gamma_K X_{iK} + \text{residual}$$

Using (7), we can express the estimated value of each estimated coefficient in the β vector in (6) as

$$(8)\quad E(\hat{\beta}_j^*) = \beta_j + \gamma_j(\alpha_1\beta_Q + \alpha_2\beta_D)$$

and, for the coefficient of interest here, the cross-sectional coefficient on years of work experience, the corresponding expression is

$$(9)\quad E(\hat{\beta}_{exp}^*) = \beta_{exp} + \gamma_{exp}(\alpha_1\beta_Q + \alpha_2\beta_D)$$

By using (9), it is easy to show that when vintage effects are present, the cross-sectional estimate of the effects of work experience on earnings is less than the true life cycle parameter and, moreover, that this bias is likely to be greater (more negative) for blacks than for whites. In (9), the value of β_{exp}, estimated from cross-sectional data, is equal to its true value plus a multiplicative bias term which is

[5]Since both V and X^* are nonstochastic, equation (7) can be viewed as a descriptive regression only. See Jan Kmenta, pp 391–93.

itself composed of two terms: one, γ_{exp}, reflecting the multiple regression relationship between vintage and years of work experience, and the other, $\alpha_1\beta_Q + \alpha_2\beta_D$, measuring the effects of vintage, operating through school quality and discrimination, on earnings. According to the vintage hypothesis, the vintage effects will be negative, since $\alpha_1 < 0$, $\alpha_2 > 0$, $\beta_Q > 0$, and $\beta_D < 0$. As for the first term, γ_{exp} is not only positive in a cross section (older workers have more years of work experience), but it will be approximately equal to one since vintage, measured by an individual's age, equals years of work experience plus age at entry into the labor market. As a result, the predicted bias is negative and $E(\hat{\beta}^*_{exp}) < \beta_{exp}$. If, as assumed by the vintage hypothesis, the vintage effects are stronger for blacks than for whites, then the underestimate is greater for blacks than for whites. Thus, the vintage hypothesis can fully account for the cross-sectional finding of increasing black-white earnings differences over the "life cycle." Note, however, that there is nothing in the vintage hypothesis that requires that the true life cycle coefficients for blacks and whites are necessarily equal.

Empirical tests of the vintage hypothesis have for the most part been indirect,[6] since life cycle data on individual earnings have been unavailable. The empirical work by Welch and by Smith and Welch has involved tracking various cohorts of workers as they age from one cross section to another.[7] In effect, this procedure simulates the life cycle, with the qualifications that earnings are observed at only two points in time and for two random samples of individuals, rather than for the same individuals. The basic procedure in both papers has been to estimate separate earnings regressions by race and by experience cohort in the two comparison years and then compare the ratio of the estimated coefficients on years of education (black-white) for the corresponding cohort in the two cross sections (for example, the cohort with one to ten years of work experience in 1960 is compared to the 1970 cohort with eleven to twenty years of experience). In the earlier paper, Welch found evidence of vintage effects: although the ratio of schooling coefficients declined monotonically with experience in each cross section, there was little or no within-cohort deterioration in the relative schooling coefficients across the cross sections. In their 1977 paper, however, Smith and Welch report that "the pattern emerging within cohorts for the elementary and secondary [school] segment is precisely the decline predicted by the cross-sectional relation" (p. 330).

There is also an inherent problem in this kind of comparison procedure: the results may be influenced by the specific characteristics of one or both of the observation years.[8] In this case, the possible problem is that the base year, 1959, was a high unemployment year (4.6 percent for white males and 11.5 percent for black males), while both of the comparison years, 1966 and 1969, were years of relatively low unemployment. Recently, Thomas Kniesner, Arthur Padilla, and Solomon Polachek have suggested that the relative rate of return to schooling (white compared to black) for young workers is positively related to the unemployment rate. Thus, they conclude that it is possible that the observed increases in the relative returns to education which Welch reports might reflect primarily the effects of cyclical labor market activity rather than either vintage-related or life cycle effects.

III. Estimating a Model of Life Cycle Earnings

As the above discussion of the cross-sectional bias suggests, there are two possible ways in which the estimation bias may be reduced. First, one could use cross-sectional data, but specify the model properly by incorporating direct measurements of vintage

[6]For attempts to test the vintage hypothesis directly by incorporating school quality measures (usually $/pupil by county or state) into a cross-sectional earnings function, see Charles Link and Edward Ratledge; Link, Ratledge, and Kenneth Lewis, John Akin and Irwin Garfinkel

[7]Welch compared cohorts from the 1960 Census and the 1967 *SEO*. Smith and Welch used the 1960 Census and the 1970 Census.

[8]It is not possible to control for these period effects in a cross section, since the effects are, presumably, uniform within each cross section.

effects. Available vintage measures are, however, crude at best: there are no appropriate summary measures of discrimination, and although some information on school quality is available, it is usually limited to per capita expenditures in highly aggregative form and measured across geographical units in varying ways.

Alternatively, one could use longitudinal data on the earnings of a set of individuals, all of whom are of the same vintage; in that case γ_{exp} is equal to zero.[9] Because we are primarily interested in differences in the life cycle pattern of earnings between blacks and whites rather than in the life cycle earnings of any single individual, we could combine all blacks and all whites in pooled cross-section and time-series regressions. The estimated equation, then, would be of the form:

$$(10) \quad Y_{it} = \beta_0 + \beta_1 X_{it,1} + \ldots \beta_{exp} EXP_{it} + \ldots \beta_k X_{it,k} + \epsilon_{it}$$

The problem in estimating (10) is that the actual data available to economists falls far short of a complete life cycle series on individual earnings for a sufficiently large sample of workers of the same vintage. A natural procedure to increase the sample size is to modify equation (10) and combine for estimation purposes individuals belonging to a number of different vintages. The problem this pooling procedure creates is that it increases (compared to estimation for a single vintage) the size of γ_{exp} and thereby, according to (9), increases the bias in the estimation of the life cycle returns to experience.[10] In effect, pooling observations across vintages introduces into the equation some of the cross-sectional bias due to improperly specified vintage effects.

The bias due to pooling is, however, not a serious problem in this case, since it is possible to compute the bias and recover the unbiased coefficient. To see this, consider equation (9) again. The term $E(\hat{\beta}^*_{exp})$ can be estimated from a cross-section earnings regression and, since an individual's vintage can be measured by his age, it is also possible to estimate equation (7) in order to find γ_{exp}. The remaining term in (9), $\alpha_1\beta_Q + \alpha_2\beta_D$, the effects of vintage on earnings, cannot be directly estimated, but it can be approximated by regression. Note that the derivative of (9) with respect to γ_{exp} is $\partial E(\hat{\beta}^*_{exp})/\partial\gamma_{exp} = \alpha_1\beta_Q + \alpha_2\beta_D$. Since α_1, α_2, β_Q, and β_D are parameters, the derivative is a constant and its value can be estimated by performing two pairs of regressions—for example, one, cross-sectional, and the other, pooled over several vintages—and then using the two sets of estimated values of $E(\hat{\beta}^*_{exp})$ and γ_{exp} to compute the finite approximation of the derivative. It is possible, then, to extrapolate to the true value of β_{exp} by using (9) since all the terms are known.[11]

IV. Cross-Sectional and Pooled Estimates of Black-White Life Cycle Earnings

Both the cross-sectional and pooled cross-section and time-series earnings models were estimated using data from the *Panel Study of Income Dynamics*. The primary advantage of the *Panel Study* for this analysis is that it provides longitudinal information on individual labor market earnings. For this analysis, data from the first eight years of the study, covering the period 1967–1974, were used.[12] The cross-sectional model was estimated using 1967 data.

The pooled analyses focused on two cohorts of black and white males, those individuals between ages 20 and 29 in 1967 and those between 30 and 39 in 1967.[13] The correspond-

[9]Restricting each regression to a single vintage is analogous to controlling for vintage effects experimentally rather than via multiple regression.

[10]From (9), Bias $= \gamma_{exp}(\alpha_1\beta_Q + \alpha_2\beta_D)$. Thus, the bias increases (decreases) proportionally as γ_{exp} increases (decreases).

[11]The true life cycle parameter can be approximated as

$$\beta_{exp} \approx E(\hat{\beta}^*_{exp\,p}) - \gamma_{exp\,p}\left(\frac{E(\hat{\beta}^*_{exp\,p}) - E(\hat{\beta}^*_{exp\,cs})}{\gamma_{exp\,p} - \gamma_{exp\,cs}}\right)$$

where p and cs refer to estimates from the pooled and cross-sectional regressions.

[12]The cumulative sample losses for the *Panel Study* have been surprisingly small. After an 11 percent loss in the second year of interviewing, the response rate has been over 97 percent in all subsequent years.

[13]These two cohorts correspond roughly to those individuals who would have completed high school in the

ing 1967 cross-sectional cohorts covered an age range seven years longer than the pooled cohorts—20–36 and 30–46 in 1967—so that identical portions of the experience-earnings profile would be compared in the cross-sectional and pooled equations.[14] Only individuals who were in the labor force in each year—either "working now or laid off temporarily" or "looking for work unemployed"—were included in the analysis.[15] This requirement eliminated individuals who were retired, disabled, or in school at any time during the eight years. It did not, however, necessarily exclude individuals who worked few or even zero hours in any single year; to do so might bias black-white comparisons if the incidence of long-term unemployment differed by race. Finally, individuals who were farmers and other self-employed workers in any year were also eliminated from the sample, since earnings for these two groups are notoriously difficult to measure and interpret.

The general empirical procedure involved estimating identical earnings equations for the corresponding cross-sectional and pooled cohorts. In order to provide estimates of γ_{exp}, identical cross-sectional and pooled equations were also estimated in which vintage, measured by an individual's age in 1967, was the dependent variable.

All of the earnings equations were estimated in two versions, first with experience in linear form only, and then again, adding experience squared. The test for within-cohort non-linear returns to experience was treated as simply an empirical issue, since the separate estimation of the experience coefficients across the two age cohorts already allowed for the conventional concave experience-earnings profile. Only for the younger cohort of white males was the quadratic term statistically significant, and consequently it was eliminated from the model for the older cohorts and for the younger cohort of blacks.

Since the estimation bias is directly proportional to the size of γ_{exp} (see fn. 10), the extent to which the pooled model provides improved life cycle parameter estimates depends quantitatively on the size of the reduction in γ_{exp}. The top half of Table 1 presents the cross-sectional and pooled estimates of γ_{exp}. As expected, the cross-sectional coefficients are approximately equal to one for all age and race groups, with the coefficient for the 20–29-year old whites evaluated at the mean number of years of experience. For the two pooled cohorts, the auxiliary coefficients on experience are all approximately equal to .6 (again, the quadratic is evaluated at the mean). Thus, the pooling procedure directly eliminates about 40 percent of the vintage-related bias.

The estimated cross-sectional and pooled coefficients on years of work experience from the earnings equation are summarized in the lower portion of Table 1. The dependent variable in the analysis is the natural *log* of real hourly earnings. The experience variable was computed for 1967 as age minus years of education minus 6, subject to the constraint that useful experience began no earlier than age 16. In subsequent years, years of experience was incremented depending on whether or not the individual worked 1,000 hours in the preceding year. The other independent variables were years of education, a test score measure,[16] union status, city size, region, the unemployment rate in the individual's county of residence, and a series of dummy variables for industry of employment. The equations were estimated by *OLS* regression.[17] The full

postdesegregation period (1955–66) and in the postwar, predesegregation period, respectively.

[14]Note that the cross-sectional and pooled models do not contain the same individuals. All individuals in the pooled equation are in the 1967 cross section, but the cross section includes older workers (30–36 and 40–46) not included in the pooled model and some individuals who were eliminated from the pooled model for the reasons given in the text.

[15]These are not the Bureau of Labor Statistics definitions of labor force status, but are answers to the question, "Now we would like to know about your present job—are you working now, looking for work, retired, a housewife, or what?"

[16]This variable measures the respondent's score on a thirteen-item sentence completion test, drawn from the verbal portion of the Lorge-Thorndike Intelligence Test. For further information on the test, see the *Panel Study*, pp. 367–71.

[17]Asatoshi Maeshiro has shown that in an autoregressive model which includes a trended independent variable, like experience, the familiar Cochrane-Orcutt

TABLE 1—ESTIMATED CROSS-SECTIONAL AND POOLED COEFFICIENTS ON YEARS OF WORK EXPERIENCE[a]

	White			Black	
	Number of Observations	Experience	(Experience)^2	Number of Observations	Experience
Dependent Variable: Vintage[b]					
Cross Section, 1967					
20–36	534	.9898[c]	.0007	230	.9920[c]
		(.0201)	(.0009)		(.0119)
30–46	628	.9894[c]	–	252	.9941[c]
		(.0064)			(.0099)
Pooled, 1967–74					
20–29	2,088	.4379[c]	.0074[c]	896	6192[c]
		(.0405)	(.0019)		(.0170)
30–39	2,528	.6090[c]	–	1,040	.6167[c]
		(.0096)			(.0163)
Dependent Variable: *ln* Real Wages					
Cross Section, 1967					
20–36	534	.0772[c]	– 0019[c]	230	.0175[c]
		(.0126)	(0006)		(.0060)
30–46	628	.0085[c]	–	252	–.0040
		(.0032)			(.0061)
Pooled, 1967–74					
20–29	2,088	0905[c]	– 0027[c]	896	0242[c]
		(.0088)	(.0004)		(0038)
30–39	2,528	.0113[c]	–	1,040	.0126[c]
		(.0020)			(0036)

[a]The full regression included years of education, union status, a test score measure, city size, region, county unemployment rate, and industry of employment. See Tables 3 and 4 for results of full regression

[b]Vintage is measured by an individual's age in 1967

[c]Significant at .01 level.

regression results are presented in Tables 3 and 4.

For the 20–29-year olds, the cross-sectional results show a different pattern of returns to experience for blacks and whites. First, the white experience-earnings profile is parabolic, while that for blacks is linear.[18] Over almost all of the relevant range of experience, the annual rate of increase in earnings for whites is greater than that for blacks. For example, the return to an additional year of work experience for a white worker with eight years of work experience is about 4.5 percent; for a comparable black worker, the predicted growth in earnings is only 1.75 percent annually. At twelve years of experience, the return to an additional year of experience is still almost one and a half percentage points higher (3.16 vs. 1.75 percent) for whites than for blacks.

The pooled results, however, suggest that the cross-sectional model does not accurately reflect the longitudinal pattern of earnings. This is especially true for black workers, where vintage effects are clearly evident. The returns to an additional year of experience increased substantially for black workers, from 1.75 percent in the cross section to 2.42 percent in the pooled model.[19] Compounded

transformation is likely to result in less efficient estimates than the *OLS* estimator. For this reason, and also because the pooled equation provides a biased estimate of ρ since it is still misspecified, *OLS* was used.

[18]The *F*-statistic for the inclusion of experience squared in addition to years of experience in linear form was 10.7 for whites and 1.4 for blacks. The critical value of *F* for one additional variable and approximately 200 degrees of freedom is 3.84. When the white equation was estimated without experience squared, the estimated coefficient on experience was .0375. In this form, the difference between the black and white coefficient on experience is statistically significant at the 1 percent level.

[19]In a one-tailed test of the hypothesis that the pooled coefficient is greater than the cross-sectional coefficient.

over eight years, this yields an average growth in earnings of 18 percent, compared to 13 percent in the cross-sectional earnings profile. For whites, the returns to experience change only slightly from the cross-section to the pooled model; the pooled experience-earnings profile is somewhat more concave than the cross-sectional one.[20]

The pooled model still shows some evidence of increasing life cycle earnings differences, although considerably less than indicated by the cross-sectional regressions. For the group of older black workers within the 20–29 year old cohort—those with ten or more years of work experience in 1967—the predicted eight-year rate of growth of earnings exceeded that for similar whites.[21] For the younger workers, however, earnings for whites still appeared to increase more rapidly with experience. Indeed, the pooled regression results suggest that the black-white earnings gap is almost zero at the time of entry into the labor market and then grows rapidly through the first decade of experience. This finding should be interpreted with caution, however, since, unfortunately, the sample contained relatively few blacks with less than four years of work experience in 1967.[22] More weight can be given to the results for workers with slightly more experience; for example, for workers with six years of experience in 1967, the earnings of whites increased about 33 percent compared to about 18 percent for otherwise similar black workers. As a result, the average earnings of the black workers fell from about three-quarters that of whites at six years of experience to about two-thirds after thirteen years.

For the older cohort of workers, there is again strong evidence of vintage effects. The cross-sectional results show clear differences between blacks and whites not only in the level of earnings, but also in the rate of change of earnings with experience. The real earnings of white workers increased by about 0.85 percent with each additional year of experience, while earnings for blacks actually fell by 0.40 percent per year. Although the black coefficient is not significantly different from zero at conventional levels, the difference between the black and white coefficients is statistically significant at the 10 percent level. If the individual life cycle pattern of earnings followed that inferred from the cross section, then black-white earnings differences would increase by 1.25 percent per year and by over 9 percent over the eight-year period.

In the pooled model, however, these apparent life cycle differences in the returns to experience disappear. For both blacks and whites, earnings growth outpaced that of the vintage of workers who preceded them. For whites, the vintage effect is moderate again; the pooled returns to an additional year of work experience are a quarter of a percentage point higher than in the cross-sectional model.[23] For blacks, the vintage effects are extremely strong, with each additional year of experience yielding a 1.26 percent increase in real earnings rather than the slight decline predicted by the cross section. The difference between the cross-sectional and pooled estimates for black workers is statistically significant at the 5 percent level and there is now no statistically significant difference between the estimated black and white returns to experience.

The estimated coefficients on experience are, as noted above, biased even in the pooled model, since the correlation between vintage

the null hypothesis could be rejected at about the 80 percent level of confidence

[20]When the pooled model was estimated without experience squared, the coefficient on experience fell to .0341, compared to .0375 in the cross-sectional equation. In this form, the equation suggests that vintage effects were negative for all of the whites in this age cohort.

[21]For workers with ten years of work experience as of 1967, the predicted eight-year growth rate of earnings was 18.2 percent for blacks and 15.1 percent for whites.

[22]There were only twelve blacks with four years of experience or less in 1967. This was in part a function of the way experience was defined and the age filter that was used. Experience in 1967 was defined as age minus years of education minus 6 (constrained to be no more than age minus 16) and the minimum age in the sample was 20. Thus, no high school graduate could have less than two years of work experience in 1967. The most likely candidates for very low experience in 1967 were young (20–22) high school graduates or slightly older (23–25) college graduates. It turned out that there were very few blacks in either category.

[23]The difference between the two coefficients is not statistically significant at conventional levels.

TABLE 2—CROSS-SECTIONAL, POOLED, AND EXTRAPOLATED COEFFICIENTS ON YEARS OF WORK EXPERIENCE FOR BLACK AND WHITE WORKING MALES, 1967 SAMPLE[a]

	Age 20–29 with:			
	Four Years Work Experience	Eight Years Work Experience	Twelve Years Work Experience	Age 30–39
White				
Cross Section	.0620	.0468	.0316	.0085
Pooled	.0689	.0473	.0257	.0113
Extrapolated	.0757	.0479	.0164	.0158
Black				
Cross Section	.0175	.0175	.0175	−.0040[b]
Pooled	0242	0242	.0242	.0126
Extrapolated	.0353	0353	.0353	0396

[a]The full regression included years of education, union status, a test score measure, city size, region, county unemployment rate, and industry of employment
[b]Not significantly different from zero at .01 level.

and years of work experience (γ_{exp}) is not equal to zero. It is, however, possible to extrapolate the true life cycle returns to experience by comparing the estimated coefficients on experience in both the cross-sectional and pooled earnings equations and in the corresponding auxiliary equations (see fn. 11). Because the extrapolation procedure is based on some strong assumptions—namely, that the cross-sectional model is correctly specified except for the omitted vintage effects and that the earnings model is itself stable over time except for changes in the returns to experience—it is probably prudent to think of the true life cycle parameter as being bracketed by the pooled and extrapolated estimates.

The extrapolated coefficients are presented in Table 2, along with the corresponding cross-sectional and pooled estimates. For the younger cohort, the returns to experience are evaluated at various levels of experience to allow for the non-linear returns to experience for whites. For the blacks in this cohort, the returns are constant at all levels of experience, since the experience term was entered in linear rather than quadratic form.

In general, since vintage effects were stronger for blacks than for whites, the effect of extrapolation is to narrow the black-white differences in returns to experience still further. For the younger workers in the 20–29 age cohort, the remaining differences are still sizeable, but they narrow rapidly with experience. For workers with eight years of experience in 1967, the apparent cross-sectional difference of almost 3 percent per year fell to about 2.5 percent in the pooled model and finally to about 1.25 percent when the coefficients were extrapolated. For the older workers in this cohort—for example, those with twelve years of experience in 1967—the extrapolated black returns are greater than the returns to experience for white workers This is also true for the entire 30–39 year old cohort, where the extrapolated black coefficient is over 2 percentage points greater than that for whites. Assuming that the pooled and extrapolated estimates bracket the true parameter value, the results suggest that differences in earnings growth did exist in the first eight to ten years of work, but that thereafter, earnings differences were maintained or even reduced. For both cohorts, the pooled and extrapolated results indicate a far more optimistic life cycle situation than would be inferred from cross-sectional results.

V. Summary

The comparison of the cross-sectional and pooled results provide general support for the vintage explanation of the increasing black-

TABLE 3—1967 CROSS-SECTION AND 1967–74 POOLED EARNINGS FUNCTIONS FOR COHORT OF YOUNG BLACK AND WHITE MALES

Variable	Cross Section (Age 20–36 in 1967)		Pooled (Age 20–29 in 1967)	
	White	Black	White	Black
Experience	.0772[a]	.0175[a]	.0905[a]	.0242[a]
	(.0126)	(.0060)	(.0088)	(.0038)
Experience²	–.0019[a]	–	–.0027[a]	–
	(.0006)		(.0004)	
South	–.0339	–.0384	–.0087	–.0625[b]
	(.0373)	(.0740)	(.0189)	(.0367)
City Size	.0002[a]	.0005[a]	.0002[a]	.0004[a]
	(.0001)	(.0001)	(.0001)	(.0001)
Union	.1581[a]	.1951[a]	.1104[a]	.1197[a]
	(.0394)	(.0630)	(.0200)	(.0287)
County Unemployment Rate	.0162[b]	.0110	.0116[a]	–.0077
	(.0095)	(.0271)	(.0035)	(.0060)
Education	.1000[a]	.0430[a]	.1020[a]	.0480[a]
	(.0800)	(.0132)	(.0045)	(.0068)
Test Score	.0095	.0086	.0044	.0189[a]
	(.0096)	(.0128)	(.0054)	(.0069)
Industry[c]				
Agriculture and Mining	–.1035	–.0947	–.1514[a]	–.2596[a]
	(.1109)	(.1351)	(.0771)	(.0743)
Manufacturing, Durables	.1616[a]	.1774[a]	.0827[a]	–.0090
	(.0652)	(.0950)	(.0340)	(.0464)
Construction, Communications, and Transportation	.1236[a]	.1308	.0084	.1173[a]
	(.0651)	(.0910)	(.0342)	(.0465)
Trade and Finance	.0031	–.0321	–.0591	–.0938[b]
	(.0684)	(.1094)	(.0354)	(.0514)
General Services	.0535	.1814	.0324	–.1100[b]
	(.0771)	(.1250)	(.0413)	(.0594)
Professional Services	.0592	.2425[a]	–.1737[a]	.1571[a]
	(.0737)	(.1246)	(.0396)	(.0639)
Government	.0862	–.0476	.0149	.0097
	(.0799)	(.1274)	(.0417)	(.0611)
Constant	1.0272	–.4576	–.8331	–.2093
N	534	230	2008	896
$\bar{R}^2$	.363	.348	.288	.309

[a]Significant at 95 percent level of confidence
[b]Significant at 90 percent level of confidence
[c]Omitted category is nondurable goods manufacturing.

hite earnings differentials found in previous ross-sectional studies. Vintage effects were ound to be especially strong for black workrs. Although the cross-sectional regressions ndicated that black-white earnings differenials did tend to increase over the life cycle, hese experience-related differences were not ound—or were to be much weaker—in the ooled analyses. When the pooled estimates vere extrapolated to their true life cycle alues, black-white differences were narrowed even further. Indeed, for workers who were between ages 30 and 39 in 1967, black earnings grew more rapidly through 1974 than did the earnings of white workers. This was also true for the older portion of the 20–29-year old cohort. Since it is difficult to imagine that the normal operation of the labor market actually favors blacks over whites, it is possible that this finding reflects the effects of compensatory government labor market programs.

TABLE 4—1967 CROSS-SECTION AND 1967–74 POOLED EARNINGS FUNCTIONS FOR COHORT OF MIDDLE-AGED BLACK AND WHITE MALES

Variable	Cross Section (Age 30–46 in 1967)		Pooled (Age 30–39 in 1967)	
	White	Black	White	Black
Experience	.0085	−.0040	.0113[a]	.0126[a]
	(0032)	(.0061)	(.0020)	(.0036)
South	−.0936[a]	−.0277	−.0737[a]	.0138
	(.0353)	(.0807)	(.0178)	(.0343)
City Size	.0004[a]	.0005[a]	.0004[a]	.0006[a]
	(.0001)	(.0001)	(0001)	(.0001)
Union	1068[a]	.2528[a]	.0801[a]	.1820[a]
	(0357)	(.0760)	(.0168)	(.0294)
County Unemployment Rate	.0035	.0225	.0057[a]	.0159[a]
	(.0090)	(.0242)	(.0030)	(.0053)
Education	.0722[a]	.0363[a]	.0732[a]	.0356[a]
	(0068)	(0137)	(.0036)	(.0055)
Test Score	.0304[a]	.0091	.0231[a]	.0186[a]
	(.0084)	(.0125)	(.0045)	(0052)
Industry[c]				
Agriculture and Mining	−.0874	−.4869[a]	−.0916[b]	−.5127[a]
	(0994)	(.1449)	(0558)	(0651)
Manufacturing, Durables	.0465	1572	.0871[a]	.1149[a]
	(.0535)	(.1075)	(.0279)	(.0473)
Construction, Communications, and Transportation	.0497	1684[a]	0989[a]	.1297[a]
	(.0550)	(.0953)	(.0295)	(.0438)
Trade	−.0860	0675	−.1017[a]	−.0083
	(.0611)	(1194)	(.0342)	(.0555)
General Services	−.0837	.2400[b]	−.0044	−.0265
	(.0685)	(.1372)	(.0361)	(0560)
Professional Services	−0776	1849	0134	.0530
	(.0608)	(.1284)	(.0314)	(.0528)
Government	−.0534	2748[a]	−.0565[b]	.1261
	(.0665)	(1325)	(0334)	(.0542)
Constant	−.1818	−.1096	−.1507	−.3349
N	628	252	2528	1040
$\bar{R}^2$	.378	448	.369	.469

[a]Significant at 95 percent level of confidence
[b]Significant at 90 percent level of confidence.
[c]Omitted category is nondurable goods manufacturing.

A prominent exception to the more optimistic life cycle results are those for black workers in the first five to ten years after entering the labor market. Over these years, the earnings of young black workers grew much less rapidly than the earnings of white workers. Since earnings do not diverge significantly thereafter, this finding suggests that many of the problems which confront blacks in the labor market may be encountered at the time of entry into the labor market.

From the standpoint of government policy, two findings stand out. One is positive, namely the apparent higher pooled and extrapolated returns to experience for black workers than for white workers in the 30-39-year old cohort. This finding suggests tha the effects of prior labor market discrimination may not be permanent and that experience-earnings profiles are not immutabl determined in the first few years of experience. The other important finding for government policy is the large difference in earning growth which existed for the younger cohor of black and white workers. Problems at th time of entry into the labor market or shortl

thereafter do not lend themselves easily to government solutions. In addition to the kind of discrimination implied by theories of labor market segmentation, there may be differences between young blacks and whites in their access to information about job opportunities or in job search. In the light of the research presented here, it would seem to be insufficient for government policy to focus exclusively on the educational system as a means for reducing black-white earnings differences.

REFERENCES

J. S. Akin and I. Garfinkel, "School Expenditures and the Economic Returns to Schooling," *J. Hum. Resources*, Fall 1977, *12*, 460–81.

K. J. Arrow, "Models of Job Discrimination" and "Some Mathematical Models of Race in the Labor Market," in Anthony Pascal, ed., *Racial Discrimination in Economic Life*, Lexington 1972.

Gary Becker, *Human Capital*, New York 1964.

Y. Ben-Porath, "The Production of Human Capital and the Life-Cycle of Earnings," *J. Polit. Econ.*, Aug. 1967, *75*, 352–65.

A. Blinder, "Wage Discrimination: Reduced Form and Structural Estimates," *J. Hum. Resources*, Fall 1973, *7*, 436–55.

——— **and Y. Weiss,** "Human Capital and Labor Supply: A Synthesis," *J. Polit. Econ.*, June 1976, *84*, 449–72.

Peter B. Doeringer and Michael J. Piore, *Internal Labor Markets and Manpower Analysis*, Lexington 1971.

R. E. Hall, "Wages, Income, and Hours of Work in the U.S. Labor Force," in Glen C. Cain and Harold Watts, eds., *Income Maintenance and Labor Supply*, Chicago 1973.

Giora Hanoch, "An Economic Analysis of Earnings and Schooling," *J. Hum. Resources*, Summer 1967, *2*, 319–29.

J. Heckman, "A Life-Cycle Model of Earnings, Learning, and Consumption," *J. Polit. Econ.*, Aug. 1976, *84*, S11–S44.

Jan Kmenta, *Elements of Econometrics*, New York 1971.

T. Kniesner, A. Padilla, and S. Polachek, "The Rate of Return to Schooling and the Business Cycle," *J. Hum. Resources*, Spring 1978, *13*, 264–77.

C. Link and E. Ratledge, "Social Returns to Quality and Quantity of Education: A Further Statement," *J. Hum. Resources*, Winter 1975, *10*, 78–89.

———, ———, **and K. Lewis,** "Black-White Differences in Returns to Schooling: Some New Evidence," *Amer. Econ. Rev.*, Mar. 1976, *66*, 221–23.

A. Maeshiro, "Autoregressive Transformation, Trended Independent Variables, and Autocorrelated Disturbance Terms," *Rev. Econ. Statist.*, Nov. 1976, *58*, 497–500.

H. E. Ryder, F. P. Stafford, and P. E. Stephan, "Labor, Leisure, and Training Over the Life-Cycle," *Int. Econ. Rev.*, Oct. 1976, *17*, 651–74.

S. Rosen, "Learning and Experience in the Labor Market," *J. Hum. Resources*, Summer 1972, *7*, 326–42.

J. Smith and F. Welch, "Black-White Male Wage Ratios: 1960–1970," *Amer. Econ. Rev.*, June 1977, *67*, 323–38.

F. Welch, (1973a) "Black-White Differences in Returns to Schooling," *Amer. Econ. Rev.*, Dec. 1973, *63*, 893-907.

———, (1973b) "Education and Racial Discrimination," in Orley Ashenfelter and Albert Rees, eds., *Discrimination in Labor Markets*, Princeton 1973.

———, "Human Capital Theory: Education, Discrimination, and Life-Cycles," *Amer. Econ. Rev. Proc.*, May 1975, *65*, 63–75.

M. Zeman, "A Quantitative Analysis of White-Nonwhite Income Differentials," unpublished doctoral dissertation, Univ. Chicago 1955.

Survey Research Center ISR, (*OEO*) *Panel Study of Income Dynamics*: *Study Design, Procedures, Available Data*, 1968–72 Interviewing Years, Vol. I, Ann Arbor 1972.

Efficient Wage Bargains under Uncertain Supply and Demand

By Robert E. Hall and David M. Lilien*

Much recent thought has been devoted to the macro-economic importance of the existence of wage contracts (see Stanley Fischer; Edmund Phelps and John Taylor; Robert Barro) and to the deeper problem of the nature of the optimal employment bargain (see Costas Azariadis, 1975, 1978; Martin Baily; Donald Gordon; Herschel Grossman, 1976, 1977; Phelps and Guillermo Calvo; and Phelps). Still, some puzzling features of the most conspicuous form of wage bargaining, that done formally by employers and labor unions, deserve further theoretical attention. Among these important features are:

1. Collective bargaining agreements are rarely contingent on outside events even though the parties have very imperfect knowledge of prospective economic conditions during the period of the contract. The only important exception is the indexing of wages to the cost of living.

2. Employers are permitted wide discretion in determining the level of employment when demand shifts unexpectedly. As employment varies, total compensation varies according to a formula established in the agreement.

3. Agreements are not permanent but are renegotiated on a regular cycle.

4. In the process of renegotiation, the current state of demand has little impact on the new wage schedule. On the other hand, current wages in other industries have an important influence. This feature especially has been denied or ignored by economic theorists even though it is a prominent part of the thinking of labor economists on wage determination.

I. The Employment Bargain

Since Wassily Leontief's classic paper, it has been known that employers and workers must agree on more than just an hourly wage. In bargaining under certainty, there is a presumption that the bargain establishes a level of employment as well as a level of compensation. A review of this case will help to establish some concepts that are useful when bargainers are grappling with uncertainty as well. First, let the technology of the firm and the demand function for its products be jointly summarized by a revenue function $R(L)$, giving gross dollar revenue as a function of total labor input L. We assume, as seems appropriate in the context of collective bargaining, that the firm has some monopoly power. Second, let $V(L)$ be the labor union's opportunity cost of supplying that amount of labor, in the sense of foregone consumer surplus from diverting leisure to work, or foregone earnings from other employment. The term $V(L)$ can be thought of as the minimum offer that the union would ever accept to supply L.[1] We do not pursue the question of how the union divides L among its members, nor the way that it allocates the proceeds of the sale of L. If the union members have identical preferences and identical alternative employment opportunities and if the union allocates work and income evenly then $V(L)$ simply represents the typical union member. These results are compatible with many other views of the internal politics of unions, however.

Labor input L is one dimension of the bargaining problem. The other is dollar compensation. It is most convenient to deal

*Stanford University and the University of California-San Diego, respectively. This research was supported by the National Science Foundation. We are grateful to Jerome Culp, Avinash Dixit, and Jeffrey Harris for useful suggestions.

[1]In general, $V(L)$ should depend on wealth as well as employment. This dependence complicates the analysis without changing its character, and is deferred to the Appendix.

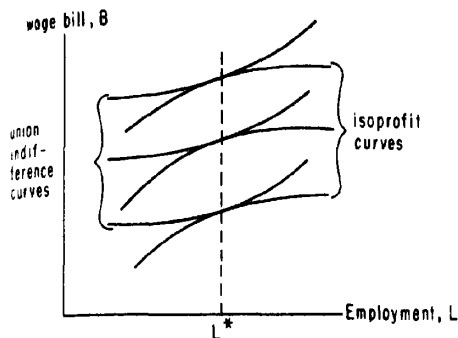



FIGURE 1

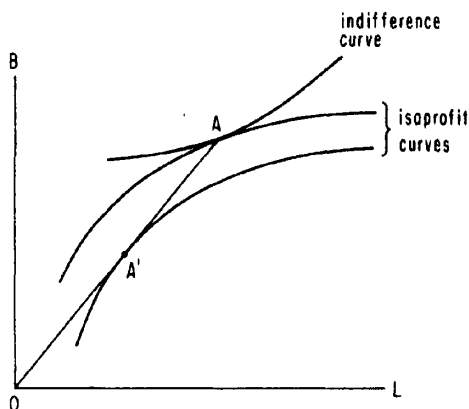



FIGURE 2

with total compensation or the wage bill, say B, rather than with the more familiar average hourly wage. The firm's concern is with profit, $\pi = R(L) - B$, and the union's concern is with income net of the opportunity cost of the time devoted to work, $Y = B - V(L)$. The bargaining problem is illustrated in Figure 1. The contract curve, or set of alternative efficient bargains, is the vertical line, $L = L^*$.[2] Each point on the line satisfies the basic efficiency requirement that the marginal revenue product of labor, $R'(L)$, equals the marginal opportunity cost of work, $V'(L)$. Assume that cooperative bargainers will always strike a bargain that is efficient. The major issue in bargaining is the determination of the wage bill, that is, the splitting of monopoly profits between firm and union. Most theoretical work on collective bargaining has concerned this very difficult issue.[3] We will avoid it altogether. Our interest is in describing efficient bargains, and in the case of complete certainty, the answer is straightforward—a bargain is efficient only if it sets a level of employment at L^*.

As Leontief pointed out, the efficient bargain cannot generally be supported by agreeing on an average wage, $w = B/L$, and then letting the firm maximize profit subject to this wage. Profit maximization sets the marginal revenue product of labor to the wage, while an efficient bargain requires that it equal the marginal opportunity cost of work. Only by coincidence would they be the same. Figure 2 illustrates the point. Suppose bargaining settled on the efficient point A. The slope of the line $0A$ is the implicit average wage, and the line itself is the set of employment levels and wage bills that the firm would choose among if it maximized profit subject to the average wage. The point of maximum profit is A', where $0A$ is tangent to an isoprofit curve, and profit here is necessarily higher than at A. So the collective bargaining agreement must specify the level of employment as well as the total compensation or average wage. Alternatively, the bargainers may use some other mechanism to ensure efficiency.

Most employers have wide latitude in setting the level of employment. It is interesting to ask whether there are contractual arrangements that achieve efficiency while permitting unilateral profit maximization by the firm. In particular, we ask if there is a way to make the wage bill a function of employment, say $B(L)$, so that maximization of profit $\pi(L) = R(L) - B(L)$ occurs at the efficient level of employment L^*. In fact, there is an infinite variety of such arrangements. Any compensation rule is satisfactory if the marginal cost of labor to the firm $B'(L)$ equals the marginal opportunity cost of labor

[2]The fact that the contract curve is vertical is a consequence of our assumption that the marginal opportunity cost of labor is independent of wealth and so is unaffected by the level of compensation. Again, see the Appendix for the more general case.

[3]George de Menil presents a complete discussion with many references.

$V'(L)$ at the efficient level of employment L^*. To avoid the possibility that this point is a minimum instead of a maximum, and to eliminate multiple local maxima, we will also require that the marginal cost of labor $B'(L)$ be an increasing function of L. Then as an alternative to stating the bargain as a particular level of employment and a particular amount of compensation, bargainers may arrive at a compensation rule $B(L)$ instead, and let firms choose employment unilaterally subject to the rule. This way of expressing the outcome of bargaining seems closer to the actual practice in American collective bargaining.

II. Efficient Wage Bargains under Uncertainty

Neither party to a collective bargaining agreement has full knowledge of the economic circumstances that will prevail during the agreement. Both the demand for products and the opportunity cost of labor can change unexpectedly. Framers of agreements must anticipate the possible need to adjust the level of employment as supply and demand change. Adjustments of employment are in fact made on a weekly schedule in most American industries.

Suppose that shifts in demand can be characterized by a single random variable x, so that the revenue function is $R(L, x)$. Though we do not make any assumptions about the probability distribution of x, it is not necessarily a variable that moves around a normal level—it may drift upward or downward. On the supply side, hypothesize a similar random shift z in the opportunity cost of employment $V(L, z)$. Again, z captures general changes in the cost of living and in wages elsewhere in the economy, as well as conditions in labor markets close to the firm where union members might seek alternative employment.

For particular values of x and z, the efficient level of employment is still defined by the equality of the marginal revenue product of labor and the marginal opportunity cost of labor:

$$(1) \qquad \frac{\partial R(L, x)}{\partial L} = \frac{\partial V(L, z)}{\partial L}$$

This defines an efficient level of employment as a function of x and z: $L^*(x, z)$. The problem that concerns us in the remainder of this paper is contractual mechanisms that permit the level of employment to fluctuate exactly, or at least approximately, according to this function. Throughout, it is assumed that it is physically possible to adjust employment as supply and demand shift, so the relevant concept of efficiency is *ex post*. Because we seek contracts that are efficient in this sense no matter what random shifts have occurred, we make no assumptions about the probability distributions of x and z, nor do we assume that these distributions are known to the bargaining parties.

One class of contracts achieves efficiency *ex post* exactly for any values of x and z. In these contracts, the level of employment is made contingent upon the values of the random shifts. The employment part of the bargain is then a function of x and z, and the bargain is efficient only if the function is $L^*(x, z)$. As in the case of certainty, the hard part of the bargaining process is settling upon a distribution of monopoly profits. Now this must be done separately for each possible contingency by establishing a contingent wage bill $B(x, z)$.

Important practical obstacles limit the scope of contingent contracts. One of these is the need to measure the variables describing the contingency quickly and accurately, and to avoid costly legal disputes about the measurements. Government statistics of high quality, such as the Consumer Price Index, are acceptable for this purpose, but many other kinds of data, publicly or privately collected, probably are not. Making contracts involving millions or billions of dollars contingent on a particular statistic obviously puts immense strain on the collectors of the statistic. The second practical obstacle is moral hazard. If the contract is contingent upon variables that can be influenced by one of the parties to the contract, they will face an incentive to change the variable accordingly.[4]

[4]These limitations on contingent labor contracts, especially moral hazard, are discussed by Phelps and Calvo, and by Phelps. In this paper, we do not present an extended justification for the assumption that contracts

The existence of moral hazard does not preclude the use of contingent contracts. It may simply be accepted as an inevitable source of inefficiency, as for example in the case of automobile insurance where insured drivers presumably are not quite as careful as they should be.

As an extreme case of the problem of moral hazard, suppose that there are no genuine shifts in demand, but the bargainers agree upon a contingent contract providing employment, $L^*(x, z)$, and compensation, $B(x, z)$. Management can in fact announce any value of x they please. Then the contract effectively establishes a marginal cost of labor which is the change in compensation per unit change in employment, $(\partial B/\partial x)/(\partial L^*/\partial x)$. To maximize profit, management will choose the x that equates the marginal revenue product of labor to the implicit marginal cost of labor. Unless this has been foreseen by the bargainers, the resulting level of employment will not be efficient. For example, suppose that the compensation side of the bargain is a guaranteed annual wage, independent of demand and supply: $B(x, z) = B_o$, a constant. Then management can cheat by driving the marginal revenue product of labor down to the implicit cost of labor, namely zero, by a suitable overstatement of the level of demand. Conversely, if the compensation rule were quite sensitive to demand, management would have an incentive to understate demand in order to reduce labor input.

Cheating of this kind does not seem to be an important problem in American labor relations. Rather, both parties agree from the beginning that management is free to maximize profit under the terms of the contract. There are no outside demand variables, supposedly beyond the control of the firm, upon which the collective bargaining agreement is contingent.

Similar considerations limit the usefulness of contingent bargains on the supply side. A reliable government cost-of-living index is published in the United States, and many agreements are contingent upon it. Much less reliable but highly relevant data on hourly wages are also published, but we are unaware of any formal contingencies based on wages. Many of the same obstacles limit the use of outside variables to measure shifts in the opportunity cost of labor as limit the use of demand variables. In the rest of this paper, therefore, we assume that contracts cannot simply specify employment as a function $L^*(x, z)$ and compensation as $B(x, z)$. More subtle procedures for achieving efficiency are required.

III. Efficient Contracts that are not Contingent on Outside Variables

Bargainers who are wary of making agreements contingent upon outside measures of demand and supply nonetheless have a powerful tool at their disposal for achieving efficiency *ex post*: They can make the level of compensation contingent upon the actual level of employment.[5] Of all the data that might be supplied by the firm to the union, the level of employment is most easily verified by the union. However, since both parties have a direct influence over the amount of employment, bargainers presumably must accept the constraint that both management and labor will separately try to maximize well-being by adjusting the level of employment.[6] This situation would breed a tension of its own unless the provisions of the contract are such as to make both parties agree on the desirable level of employment *ex post*.[7]

are not contingent, but rather pursue the implications of the assumption. In our view, it is still an unsettled question why contingencies are so rare.

[5]Again, this type of contingency is discussed by Calvo and Phelps.

[6]An interesting discussion of this problem appears in Herbert Simon, ch. 11. Simon asks under what conditions labor will permit management to make a unilateral employment decision in exchange for a lump sum wage payment. He does not consider the generalization where management makes the employment decision subject to a more general compensation rule.

[7]In the usual contingent contract of the kind studied by Azariadis (1975), for example, optimization is prohibited. However, the limited ability of employers to enforce the terms of an employment contract against workers makes it possible for workers to default by quitting, as discussed by Grossman. Under the type of contract considered here, workers never have an economic incentive to quit.

The collective bargaining agreement we study here establishes a compensation rule $B(L)$, *not directly contingent upon the* demand and supply shifts x and z. The firm has profit, $R(L, x) - B(L)$, and the union has well-being, $B(L) - V(L, z)$. The efficient level of employment is the same $L^*(x, z)$ defined earlier. We ask: Is there a compensation rule under which the quantity of labor demanded by management derived from

$$\frac{\partial R(L, x)}{\partial L} = B'(L) \tag{2}$$

is equal to the quantity of labor supplied by the union derived from

$$\frac{\partial V(L, z)}{\partial L} = B'(L) \tag{3}$$

for all possible x and z? Under such a compensation rule, management and labor will never disagree about the desirable level of employment. In particular, there is no danger of default by either party to the contract. Furthermore, *a compensation rule that achieves agreement in this sense is automatically efficient ex post* Since $\partial R/\partial L$ and $\partial V/\partial L$ are both equated to the same $B'(L)$, they equal each other, and this is the definition of efficiency.

Compared to the fully contingent bargain discussed in the previous section, only one major restriction limits the ideal employment-contingent contract: The distribution of monopoly profit cannot be made separately contingent on shifts in demand and supply. The shape of $B(L)$ is determined by the efficiency requirement alone. As L shifts, the distribution of monopoly profits shifts as a by-product of the shape of the compensation rule. The bargainers retain the single most important distributional tool—they are free to determine the overall level of $B(L)$ by adding a constant to it. Efficiency depends only on the marginal rate of compensation $B'(L)$ and not on its level. But the ideal employment-contingent contract is incapable of insuring workers against fluctuations in income, unless $B(L)$ provides the appropriate amount of insurance.[8]

Under the most general conditions, the ideal employment-contingent compensation rule does not exist. A fixed relation between compensation and employment does not have enough degrees of freedom, so to speak, to achieve efficiency under all possible combinations of demand and supply shifts. This proposition is formalized and proved in the Appendix.

Ideal compensation rules do exist under more restrictive conditions. First, suppose that there are no uncertainties about the opportunity cost on the supply side, or, more realistically, all uncertainties can be eliminated through the use of a cost-of-living escalator. Then $V(L, z)$ is just $V(L)$. The ideal compensation rule satisfies

$$\frac{\partial R(L, x)}{\partial L} = B'(x), \text{ all } x, \tag{4}$$

and

$$V'(L) = B'(L)$$

The second equation suggests the nature of the ideal rule. Firms should be made to internalize labor's opportunity costs by paying a marginal cost of labor $B'(L)$, which always equals labor's marginal opportunity cost $V'(L)$. In other words, the compensation rule has the form

$$B(L) = B_o + V(L) \tag{5}$$

where B_o is a fixed cash payment to labor which is independent of the amount of work and is chosen by the negotiators in the light of purely distributional considerations. Again, we have nothing to say about B_o, but any value of B_o is consistent with efficiency. Under this compensation rule, the well-being of the union is

$$\begin{aligned} B(L) - V(L) &= B_o + V(L) - V(L) \\ &= B_o \end{aligned} \tag{6}$$

[8]Here we depart from Calvo and Phelps. They define the preferred contract as the optimal compromise between the goals of efficiency and insurance. The compromise is necessary because of the limited power of the employment contingency. However, under the assumptions we use in the body of the paper, the efficient contract also provides optimal insurance. The effect of modifying the assumptions is discussed in the Appendix. We will return briefly to the insurance issue at the end of the paper.

a constant independent of L. Labor never disagrees with management about L because it is *indifferent* among all levels of employment. Rather than a guaranteed annual wage, labor has a guaranteed annual level of well-being. Management alone knows the true level of demand, measured by x, and makes a *unilateral* decision about employment. The level chosen is efficient because of the internalization of labor's opportunity costs by management. The absence of contractual limitations on employer's rights to vary employment is one of the four major features of American collective bargaining that we set out to explain. In that respect, contracts where the compensation rule embodies the opportunity cost of labor's time may be realistic.[9]

At the opposite extreme, suppose that demand is known in advance, so that the revenue function is just $R(L)$, but random shifts in the supply side make labor's opportunity cost depend on z: $V(L, z)$. An ideal compensation rule then satisfies

$$R'(L) = B'(L) \tag{7}$$

$$\frac{\partial V(L, z)}{\partial L} = B'(L) \tag{8}$$

Again, such a rule exists for this case, but now it has an entirely unfamiliar flavor. Compensation must be total revenue less a distributive constant:

$$B(L) = R(L) - B_o \tag{9}$$

Profit is

$$\pi(L) = R(L) - B(L) = B_o \tag{10}$$

so management is indifferent among all levels of employment. Labor internalizes the production function and product demand functions and unilaterally determines employment to maximize its "profit," $R(L) - V(L, z) - B_o$. No collective bargaining agreement known to us provides that the union makes the employment decision and management acquiesces passively, even though this kind of contract seems a theoretical possibility.

The striking difference between the two cases makes it clear why an ideal employment-contingent contract does not exist when both demand and supply shift randomly. To achieve efficiency in the face of demand shifts, the marginal rate of compensation $B'(L)$ must *increase* with L, as it reflects the increasing marginal opportunity cost of labor. But to achieve efficiency in the face of supply shifts, $B'(L)$ must decrease with L, because the marginal revenue product of labor is a decreasing function. The two requirements are contradictory. The contradiction is also apparent between the administrative arrangements that support the two kinds of contracts. One or the other party, but not both, must set employment unilaterally.

One other special case will prove useful for the ensuing discussion. Suppose that both supply and demand shift, but that the two shifts are related. Specifically, let $z = f(x)$. The conditions for an ideal compensation rule are

$$\frac{\partial R(L, x)}{\partial L} = B'(L) \tag{11}$$

$$\frac{\partial V(L, f(x))}{\partial L} = B'(L) \tag{12}$$

These define the efficient level of employment as a function $L^*(x)$ of the random shift. The most interesting and relevant case is where the shift raises the marginal revenue product of labor and raises the marginal opportunity cost of labor, but the former predominates so the efficient level of employment rises when x rises. For example, this would describe a cyclically sensitive industry where x is a measure of aggregate demand.

In this case the ideal compensation rule exists but does not have a closed mathematical form. On the margin, it can be described in the following way. Let $x(L)$ be the value of the random shift for which L is the efficient level of employment (the inverse of $L^*(x)$).

[9]Under the fixed wage contract that is studied in the pioneering work of Baily and Azariadis, the level of employment is made contingent on an outside measure of demand. Lower levels of employment are desirable under that type of contract when demand is low if labor values its time. The novel feature here and in Calvo and Phelps is to make management responsible for setting employment and to use profit maximization *ex post* as the mechanism for achieving the efficient level of employment.

Then

$$(13)\quad B'(L) = \frac{\partial V(L, f(x(L)))}{\partial L}$$

or

$$(14)\quad B(L) = B_o + \int_o^L \frac{\partial V(L, f(x(L)))}{\partial L} dL$$

where B_o is a purely distributional parameter, as before. The principle remains that marginal compensation should equal the marginal opportunity cost of labor, but now the opportunity cost must be evaluated at the value of the supply shift z, corresponding to the demand shift x. Under our assumptions, the dependence of the supply shift on the demand shift has a particular implication about the shape of $B(L)$: Marginal compensation rises more sharply with L than does the marginal opportunity cost of labor.

Under the ideal compensation rule, management determines employment unilaterally by setting L to satisfy profit maximization.

$$(15)\quad \frac{\partial R(L, x)}{\partial L} = B'(L)$$

It is interesting to compare the response of employment to a change in x in this case to the response when supply is unaffected by x. In the present case, where the compensation rule takes account of related shifts in supply, the steeper marginal compensation schedule acts to limit employment fluctuations. The logic is straightforward—if demand shocks are accompanied by offsetting supply shocks, then the efficient change in employment is smaller than it would be if supply were unrelated. An important special case of this proposition has been pointed out by Barro: Purely monetary disturbances generate exactly offsetting shifts in supply and demand and leave the efficient level of employment unchanged. If collective bargaining achieves efficiency *ex post*, then monetary disturbances should have no effect on employment even within the span of labor contracts. This is a major criticism of existing contract theory which frequently claims to explain the apparent sensitivity of employment to monetary disturbances. We will return to this point at the conclusion of the paper.

IV. Approximately Efficient Contracts

When both demand and supply are uncertain, ideal employment-contingent compensation rules do not exist. Collective bargainers seem to deal with this problem in the following way: Approximately efficient contracts are adopted that preserve efficiency in the face of the most likely source of large fluctuations. The contracts have a finite duration, almost always three years.[10] At the time of renegotiation, events that have caused the previous contract to become inefficient can be taken into account. An important feature of this approach to collective bargaining is the irrelevance at renegotiation of events which have already been accommodated within the existing contract. To the extent that the previous contract was successful in establishing a framework for adapting to the major contingencies, only minor contingencies will remain to be accommodated. To the naive observer, then, collective bargainers will appear to be preoccupied with trivial issues and to ignore what is most important. In particular, we believe this accounts for the strong belief among many economists that collective bargainers are perverse and irrational in failing to let the wage respond appropriately to changes in demand. We will argue that, to a first approximation, the state of demand *should* be irrelevant to rational collective bargainers.

We begin with the hypothesis that the stochastic natures of unexpected fluctuations in demand and in supply are rather different. The demand for the products of a single firm fluctuates with large amplitude, and some, but not all, of the fluctuations are transitory. Many demand shifts are unrelated to fluctuations in aggregate demand. On the other hand, fluctuations in the opportunity cost of labor are relatively small and are closely related to developments elsewhere in the economy, especially unexpected changes in

[10]In the United States, there are legal restrictions on the enforcement of contracts lasting more than three years.

average wages and prices and in the availability of work in neighboring markets. When these changes occur, they are frequently permanent. Some of the fluctuations in demand and supply are related—for example, an increase in aggregate demand raises both the demand for the firm's product and the opportunity cost of the labor it uses. But the correlation on this account is weak. Most fluctuations in demand are unrelated to aggregate demand, and although these fluctuations are invisible in aggregate statistics, they are crucial in designing an efficient labor contract.

Collective bargainers choose between two very different types of contracts in designing an approximately efficient bargain. They may establish a compensation rule based on expectations about the opportunity cost of labor, and let management choose the level of employment unilaterally in view of their perceptions of demand. Alternatively, the compensation rule may be based on expectations of the level of product demand, and labor given the task of determining the level of employment unilaterally in response to supply conditions. The first type is efficient for demand fluctuations and the second for supply fluctuations. Neither is ideal. There is no way to compromise between them without creating some kind of joint decision-making body to determine the level of employment.

Under our hypotheses about the nature of fluctuations in demand and supply, it is apparent that the first type of contract will be adopted by the bargainers. The dominant source of potential inefficiency over the life of a three-year contract is demand fluctuations. But supply fluctuations need not be completely ignored. To the extent that supply and demand shifts are correlated, management can be induced to treat the demand shift as a signal of a supply shift as well. Suppose that the supply variable z can be decomposed into a part that is predictable from the demand variable, say $f(x)$, and an unpredictable residual $\tilde{z}$:

$$z = f(x) + \tilde{z} \tag{16}$$

Under these conditions, an approximately efficient contract can be drawn along the lines suggested in the previous section. The marginal rate of compensation should be

$$B'(L) = \frac{\partial V(L, f(x(L)))}{\partial L} \tag{17}$$

The resulting contract will be exactly efficient in the face of any demand shift that is accompanied by the typical supply shift. Again, if the efficient level of employment rises with rising demand and if supply shifts away from the firm, typically, when demand rises, then the compensation rule based on this principle will result in more stable employment than will a rule based only on demand shifts. This stabilization is achieved by making it financially unattractive to management to make large adjustments in employment, not by prohibiting them.

V. Institutional Aspects

We have been deliberately vague up to this point about the mechanisms used to achieve the variations in employment that are needed to preserve efficiency. Members of effective labor unions are strongly attached to their jobs because the union obtains a share of the monopoly profit for them, so it is implausible that variations in our L are achieved by changing the number of job holders. Our apparatus is based on the contrary assumption that variations in L are achieved by changes in the amount of work done by a fixed group of union members. Part of these variations correspond to the well-understood process of changing weekly hours of work in response to shifts in demand. But data on unionized industries show equally important fluctuations in employment. Why would a powerful union let some of its members lose their jobs? If decreases in L are brought about by discharging union members, and subsequent increases in L are achieved by hiring new workers from the labor market at large, then the union is failing in its task of protecting the interests of its members.

The answer to this puzzle seems to lie in the hypothesis that relatively few of the people who stop working when demand falls actually lose their jobs. Reductions in employment are achieved by layoffs, and workers who are laid off are generally recalled after a matter of

weeks or months and resume their old jobs (see Martin Feldstein and Lilien). This process takes place even at normal levels of demand, so reduced layoff rates and faster recall are methods for *increasing* L when demand is strong, as well. In other words, an important method for varying L is to change the annual weeks of work of a fixed labor force. In some industries, notably automobiles, this is accomplished by laying off the workers in entire plants for single weeks whenever output threatens to exceed demand. In other industries, layoffs are more selective and last longer for each individual.

The compensation rules suggested by our analysis make the marginal rate of compensation a fairly sensitive function of the level of employment. Management must pay a sharp premium to obtain unusually high levels of labor input, but escape very little of the obligation to pay compensation when they use low volumes of labor. The institutional arrangements that impose the disincentive to excess employment are, first, overtime premiums that require that marginal excess hours be paid at 50 percent above the average wage,[11] and, second, provisions that require the promotion of existing workers and limit new hires to entry levels. With respect to low levels of demand, the most obvious arrangement to depress the marginal rate of compensation is the supplementary unemployment benefit. The part of state-administered unemployment insurance that is experience rated has the same effect. Further, some collective bargaining agreements, notably in the steel industry, provide that workers who are moved downward in the job ladder as a result of "bumping" are protected against the corresponding reduction in pay.

The evidence on compensation rules in American collective bargaining seems consistent with a marginal opportunity cost of labor schedule that rises fairly rapidly in the range of full-time work, though we emphasize that our theory predicts a more steeply rising marginal rate of compensation than marginal opportunity cost when demand and supply fluctuations are positively correlated. Empirical evidence on labor supply shows extreme sensitivity of the marginal opportunity cost of all labor supplied by an individual, in that estimated labor supply functions are virtually unresponsive to wages in the case of the adult men who are the bulk of union membership. However, the opportunity cost of all labor is not the same as the opportunity cost of labor supplied to the firm signing the collective bargaining agreement, since many union members hold second jobs or are able to find temporary work during periods of layoff. For them, the marginal opportunity cost of work at the union job does not fall quickly to zero below full-time work. Similarly, the part of unemployment compensation that is not experience-rated makes time spent on layoff have a positive value to workers. All of these considerations make it desirable for the compensation rule to provide incentives to reduce employment in times of slack demand.

VI. Risk and Insurance

Efficient contracts expose workers to risk. Using the compensation schedule to make the employment-contingent contract efficient rules out the alternative of smoothing the variability of income by divorcing compensation from the level of employment. In this respect, our study is the polar opposite of most other work on labor contracts. Under the assumption of convenience adopted in the body of the paper that the marginal opportunity cost of labor does not depend on wealth, it turns out that the efficient contract does not expose the union to any risk at all—the union is exactly compensated for the stochastic variation in employment arising from fluctuations in demand. But this is an artifact of our otherwise very convenient assumption. In general, the expected utility of the union could be increased by a set of insurance payments and premiums with expected value zero across the various possible levels of product demand. Were it not for the problems of moral hazard and unobservability of demand that limit the contracting process, either the firm or a private insurance company could sell insurance to the union. The work of Calvo and Phelps suggests that it is difficult to reach any definite conclusions

[11] Some premiums are required by law.

about the optimal contract that balances efficiency against risk aversion. Rather than pursue the insurance side of labor contracts any further here, we will limit ourselves to listing the considerations that would make the insurance side particularly important:

1. If alternative levels of employment have large effects on wealth, then the need for insurance against them will be correspondingly large. Thus, if union members have little income apart from compensation under the contract, and if the movements in employment that take place under the contract are permanent, then the optimal contract will sacrifice a considerable amount of efficiency. On the other hand, if the movements are largely transitory, the need for insurance is weak and efficiency will be the dominant consideration in drawing the contract.

2. Insurance becomes important when the marginal opportunity cost of labor is sensitive to the level of wealth, or, to put it another way, the elasticity of labor supply to wealth is strongly negative. Research on labor supply has shown quite weak income inelasticities for adult males.

3. Obviously, the demand for insurance depends on the degree of risk aversion on the part of the union.

4. Finally, the demand for insurance depends on the magnitude of the fluctuations that are likely to occur under the efficient contract. The incentive to insure against small fluctuations is weak.

VII. Macro-Economic Implications of Efficient Employment-Contingent Contracts

Labor contracts are of critical interest in macroeconomics because they seem to offer an explanation for the persistence of unemployment. If wages are unresponsive to general economic conditions during the three years of a contract, then it follows that wages cannot fully offset an aggregate disturbance until every contract has been renegotiated. This appears a more satisfying explanation than the principal alternative where persistence is attributed to the slow diffusion of information in the labor market.[12] But as Barro has pointed out, purely nominal aggregate shocks, such as those caused by unexpected monetary developments, have no effect on the efficient level of employment or unemployment, and framers of labor contracts ought to be able to figure out a way to avoid inefficient responses to these shocks. If all disturbances are purely nominal, then the efficient contract is extremely simple—it specifies a predetermined level of employment. The mere existence of labor contracts does not explain the persistence of unemployment satisfactorily.

Our discussion of the contracting problem at the level of the firm and the union suggests the difficulties in achieving the efficient response to an economy-wide disturbance. Employment-contingent contracts seem a workable solution to the collective bargaining problem at this level, but they embody a strong limitation on the response to aggregate disturbances. Only the part of the supply side of such a disturbance that can be predicted from the firm's own demand can be offset under the terms of the contract. If most fluctuations in demand at the level of the firm are unrelated to aggregate demand and thus are unrelated to changes in the opportunity cost of the labor supplied to the firm, then demand is a very noisy signal and the degree of offset (measured by our $f(x)$) is relatively weak. As perceived at the level of the firm, where recessions are just one of many sources of fluctuations, the inefficiency is fairly small, but in the aggregate, where only aggregate fluctuations are visible, this limitation becomes important. A recession, in this view, is treated by firms as no more than another reduction in demand for which a reduction in employment is the efficient response. Within the framework of employment-contingent bargains, there is no way to provide employers with a separate incentive to stabilize employment in the face of an economy-wide reduction in demand.

VIII. Concluding Remarks

We conclude by summarizing the explanations provided by our theory of the four puzzling features of collective bargaining listed at the outset:

[12]For a discussion of this point, see Hall.

1. *Absence of contingencies on outside variables.* Moral hazard and imperfect measurement limit contingencies based on outside variables. Further, contingency on the level of employment is such a powerful tool that it dominates outside contingencies.

2. *Unilateral determination of employment by management.* This is the central feature of employment-contingent contracts. The compensation rule makes labor approximately indifferent among alternative levels of employment, so there is little or no disagreement with management's choice. Neither management nor labor has an important incentive to default on the contract.

3. *Periodic renegotiation of contracts.* Changes in the opportunity cost of labor not predictable from changes in demand generate cumulative inefficiencies that can be relieved only by renegotiation.

4. *Irrelevance of the current state of demand in collective bargaining.* Employment-contingent contracts provide a complete mechanism for taking account of unexpected shifts in demand. Both parties are satisfied with the current level of employment if there have been no unexpected shifts on the supply side, no matter what is the level of demand. Only supply issues are sorted out in collective bargaining. This explains the paradox of large increases in wage schedules emerging from collective bargaining that takes place at a time of depressed employment. Labor's conventional justification for these increases—that wages elsewhere have risen substantially—is exactly supported by our analysis.

Appendix: Efficient Wage Bargains when the Opportunity Cost of Labor Depends on Income

A more general description of preferences would make the opportunity cost of labor depend on the overall level of well-being u, according to a function $V(L, u, z)$. Well-being depends on income from sources other than employment of union members. Further, in accordance with the life cycle-permanent income hypothesis, u should reflect expectations of *future* earnings as well as the effect of the current contract. In the absence of demand and supply fluctuations, the basic condition for efficiency is

$$R'(L) = V_1(L, u) \tag{A1}$$

(V_1 is the derivative of V with respect to its first argument.) The contract curve is no longer a vertical line in the L-B diagram of Figure 1, but presumably slopes downward, assuming that the marginal opportunity cost of work is higher at higher levels of well-being. The efficient compensation rule is defined implicitly by

$$B'(L) = V_1(L, g(L, B(L))) \tag{A2}$$

an ordinary differential equation. Here $g(L, B)$ is the level of well-being achieved when work is L and compensation is B.

With random demand and supply shifts x and z, a fully contingent contract providing compensation $B(x, z)$ is efficient if

$$\frac{\partial R(L, x)}{\partial L} = V_1(L, g(L, B(L), z), z) \quad \text{for all } x \text{ and } z \tag{A3}$$

Then we can formalize the impossibility of an ideal compensation rule:

THEOREM: *Suppose that all of the second derivatives of R and V are continuous, that both x and z can vary within some rectangular region* $a_1 \leq x \leq a_2$ *and* $b_1 \leq z \leq b_2$, *and that the disturbances matter within this region in that* $\partial^2 R/\partial L \partial x > 0$ *and* $\partial^2 V/\partial L \partial z > 0$. *Suppose further that the marginal revenue schedule slopes downward:* $\partial^2 R/\partial L^2 < 0$ *and nonwork is a normal good* $\partial V_1/\partial L + (\partial V_1/\partial g)(\partial g/\partial L) \geq 0$. *Then there is no compensation rule for which*

$$\frac{\partial R}{\partial L} = B'(L) \tag{A4}$$

and

$$\frac{\partial V}{\partial L} = B'(L) \tag{A5}$$

for all values of x and z in the region.

PROOF:

Suppose there were such a $B(L)$. Then

(A6) $$\frac{\partial R(L, x)}{\partial L} = B'(L)$$

makes L a function of x, with $\partial L/\partial x \neq 0$.

(A7) $$V_1(L, g(L, B(L), z), z) = B'(L)$$

makes L a function of z, with $\partial L/\partial z \neq 0$.

(A8) $$\frac{\partial^2 R}{\partial L^2}\frac{\partial L}{\partial z} = B''(L)\frac{\partial L}{\partial z}$$

But since $\partial L/\partial z \neq 0$, $B''(L) = \partial^2 R/\partial L^2 < 0$.

(A9) $$B''(L)\frac{\partial L}{\partial x} = \left(\frac{\partial V_1}{\partial L} + \frac{\partial V_1}{\partial g}\left(\frac{\partial g}{\partial L} + \frac{\partial g}{\partial B}B'(L)\right)\right)\frac{\partial L}{\partial x}$$

But since $\partial L/\partial x \neq 0$,

(A10) $$B''(L) = \frac{\partial V_1}{\partial L} + \frac{\partial V_1}{\partial g}\left(\frac{\partial g}{\partial L} + \frac{\partial g}{\partial B}B'(L)\right) \gtrless 0$$

a contradiction.

REFERENCES

C. **Azariadis,** "Implicit Contracts and Underemployment Equilibria," *J. Polit. Econ.*, Dec. 1975, *83*, 1183–202.

——, "Escalator Clauses and the Allocation of Cyclical Risks," *J. Econ. Theory*, June 1978, *18*, 119–55.

M. N. **Baily,** "Wages and Employment under Uncertain Demand," *Rev. Econ. Stud.*, Jan. 1974, *41*, 37–50.

R. J. **Barro,** "Long-term Contracting, Sticky Prices, and Monetary Policy," *J. Monet. Econ.*, July 1977, *3*, 305–16.

G. A. **Calvo and E. S. Phelps,** "Employment Contingent Wage Contracts," in Karl Brunner and Allan Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference Series on Public Policy, *J. Monet. Econ.*, Suppl., 1977, 160–68.

M. S. Feldstein, "The Importance of Temporary Layoffs: An Empirical Analysis," *Brookings Papers*, Washington 1975, *3*, 725–44.

S. Fischer, "Long-term Contracts, Rational Expectations, and the Optimal Money Supply Rule," *J. Polit. Econ.*, Feb. 1977, *85*, 191–205.

D. F. Gordon, "A Neo-Classical Theory of Keynesian Unemployment," in Karl Brunner and Allan Meltzer, eds., *The Phillips Curve and Labor Markets*, Vol. 1, Carnegie-Rochester Conference Series on Public Policy, *J. Monet. Econ.*, Suppl., 1975, 65–97.

H. I. Grossman, "Risk Shifting and Reliability in Labor Markets," *Scand. J. Econ.*, No. 2, 1977, *79*, 187–209.

——, "Risk Shifting, Layoffs, and Seniority," unpublished paper, Brown Univ., Nov. 1976.

R. E. Hall, "The Rigidity of Wages and the Persistence of Unemployment," *Brookings Papers*, Washington 1975, *2*, 301–335.

W. Leontief, "The Pure Theory of the Guaranteed Annual Wage Contract," *J. Polit. Econ.*, Feb. 1946, *56*, 76–79.

D. Lilien, "The Cyclical Pattern of Temporary Layoffs in United States Manufacturing," *Rev. Econ. Statist*, forthcoming.

George de Menil, *Bargaining: Monopoly Power versus Union Power*, Cambridge, Mass. 1971.

E. S. Phelps, "Indexation Issues: A Comment on the Blinder and Fischer Papers," in Karl Brunner and Allan Meltzer, eds., *Stabilization of the Domestic and International Economy*, Vol. 5, Carnegie-Rochester Conference Series on Public Policy, *J. Monet. Econ.*, Suppl., 1977, 149–159.

—— **and J. B. Taylor,** "Stabilizing Powers of Monetary Policy under Rational Expectations," *J. Polit. Econ.*, Feb. 1977, *85*, 163–90.

Herbert A. Simon, *Models of Man*, New York 1957.

The Optimal Tradeoff between the Probability and Magnitude of Fines

By A. MITCHELL POLINSKY AND STEVEN SHAVELL*

Fines are used in a variety of situations to control activities which impose external costs.[1] Examples of such activities include polluting the air, speeding or double parking, evading taxes, littering a highway, and attempting to monopolize an industry.[2] If it were costless to "catch" or "observe" individuals (or firms) when they engage in an externality creating activity, presumably everyone would be caught and fined an amount equal to the external cost of the activity. This is simply the traditional Pigouvian tax solution. Individuals would then engage in the activity only if their private benefits exceed the external cost.

However, in most situations it is difficult or costly to catch individuals who impose external costs. If, as a result, individuals are caught with probability less than one, it is often observed that the fine could be raised to a level such that, as in the Pigouvian solution, individuals would engage in the activity only if their private gains exceed the external cost. Since this can apparently be done for any given probability of catching individuals and since it is normally costlier to catch a larger fraction of those engaging in the activity, it is frequently argued that the probability should be as low as possible. The only constraint on lowering the probability that is recognized is the inability of individuals to pay the fine; thus, the optimal fine implied by this argument equals an individual's wealth. Note that this reasoning applies regardless of the external cost of the activity.[3]

This view of the optimal tradeoff between the probability and magnitude of fines does not seem realistic. Individuals are rarely if ever fined an amount approximating their wealth, especially for activities which impose relatively small external costs.

The present paper points out an error in this view and suggests an explanation of the choice of the probability and fine which seems consistent with reality. The mistake is that the view does not properly take into account the possibility that individuals may be risk averse. The possibility of risk aversion does *not* imply that individuals cannot be induced to make the same decision about engaging in the activity as they would under the Pigouvian solution. (For any given probability, the fine can be lowered from the level at which its expected value equals the external cost to a level, reflecting risk aversion, such that only those individuals for whom the private gains exceed the external costs engage in the activity.) Rather, the error is that the view does not take into account the risk that is borne by those who do engage in the activity. This risk is present whenever those individuals have to pay a fine with a (positive) probability less than one.

*Stanford University and National Bureau of Economic Research, and Harvard University, respectively. Polinsky's research on this paper was supported by the Ford Foundation, through a grant to the Program for Basic Research in Law at Harvard Law School, and by the Russell Sage Foundation. Shavell's research was supported by the National Science Foundation (grant #SOC-76-20862). Able research assistance was provided by James Foote and helpful comments were received from Robert Pollak and Daniel Rubinfeld.

[1]Civil fines are used routinely by federal, state, and local governments. For example, the widespread and increasing use of money penalties by federal administrative agencies is documented by Harvey Goldschmid. Criminal fines are, of course, also used extensively (see Alan Samuels).

[2]Similar situations arise in a private context, for example, in monitoring the effort of employees. Although fines are not used often, other monetary incentives exist, such as promotion opportunities.

[3]This argument was apparently first made explicit by Gary Becker (pp 183–85, 191–93) and has been accepted by many others. See, for example, Kenneth Elzinga and William Breit (ch. 7), William Landes and Richard Posner (pp. 10–11), and Richard Posner (pp 164–72). Related issues have also been considered by Serge-Christophe Kolm, Gordon Rausser and Richard Howitt, Balbir Singh, T. N. Srinivasan, George Stigler, Gordon Tullock (pp. 151–68), and Laurence Weiss.

The paper is organized as follows. Section I describes the problem of maximizing social welfare when individuals choose whether to participate in an externality creating activity. Section II shows that, as would be expected, if individuals are risk neutral, the optimal probability is as low as possible and the fine is as high as possible. However, the assumption of risk neutrality seems implausible in situations in which individuals would face the risk of losing all of their wealth. Section III proves two propositions if individuals are risk averse. First, if the cost of catching individuals is sufficiently small, the optimal probability equals one. This is true because the disutility from bearing the risk of being caught and fined outweighs the potential savings from a reduction in the probability. Also, when the optimal probability is one, the optimal fine equals the private gain of those who engage in the activity. Second, if it is optimal to control the activity at all, then, regardless of how costly it is to catch individuals, it may never be optimal to catch them with a very low probability and to fine them much more than the external cost. This is true because doing so would lower utility due to risk bearing and could more than offset the benefits from controlling participation in the activity. Section IV shows that it could not be beneficial to allow individuals to reduce the risk of being fined by the purchase of insurance against fines. This is true because the government could achieve the same result by lowering the fine. Section V illustrates the main points of the paper with some hypothetical calculations of the optimal probability and fine for double parking. Section VI discusses some additional features of the problem of determining the optimal tradeoff between the probability and magnitude of fines.

Although this paper does not take considerations of justice into account, it is obvious that they are relevant to determining the desired probability and fine. For example, fining individuals far in excess of the external cost they impose on society may be thought of as unfair, or catching only a small fraction of those who impose the external costs may be seen as arbitrary. These considerations appear to complement the conclusions of this paper if individuals are risk averse.

I. The Model

Individuals, who are assumed to be identical, are faced with the choice of whether to engage in an externality creating activity. What an individual would gain from engaging in it depends on random factors but is known to him before he has to make the decision.[4] If an individual chooses to engage in the activity, there is a chance that he will be caught and fined. The government finances its efforts to catch individuals by a per capita tax. Fines collected from those who are caught are used to reduce this tax. Individuals are allowed to insure fully against the risk of bearing the external cost, but are not allowed to insure at all against the risk of being fined.

To be more specific, define the following notation.

U = function giving utility of wealth. This function is assumed to reflect risk neutrality or risk aversion.

y = initial wealth.

a, b = possible gains from engaging in the activity ($a < b$). Although the gains are treated as nonmonetary in nature, they are assumed to have monetary equivalents, a and b.

q = probability that the gain would be a ($0 < q < 1$).

n = proportion of the population who engage in the activity. The decision to engage in the activity is made by individuals and is described below.

e = external cost created whenever an individual engages in the activity. Each individual in the population is assumed to be equally likely to bear this external cost.

π = per capita insurance premium for full coverage against the external cost. The premium equals the expected external cost per capita,[5]

$$\pi = ne \tag{1}$$

f = fine collected from an individual who engages in the activity and is caught. It is

[4]For example, the gain from double parking is easily imagined to vary with circumstances.

[5]Since risk-neutral individuals would be indifferent between having the insurance and bearing the risk, it does not affect the analysis to assume that they pay the premium.

assumed that the fine does not depend on the individual's private gain because it is too costly (or impossible) for the government to determine the gain.

p = probability that an individual who engages in the activity is caught. The probability of being caught is also assumed to be independent of the individual's gain.

$c(p, \lambda)$ = per capita cost of maintaining p as the probability of catching those who engage in the activity, where $c_p > 0$;[6] λ is a shift parameter of the cost function, where $c_\lambda > 0$ and $c(p, 0) = 0$.[7]

t = per capita taxes. The government sets the tax to finance the cost of catching individuals in excess of the fine revenue collected,[8]

$$t = c(p, \lambda) - npf \tag{2}$$

An individual would engage in the activity if the expected utility of doing so—taking into account the gain and the probability of having to pay the fine—exceeds the utility of his initial wealth. Thus, an individual who would gain a would engage in the activity if

$$(1-p)U(y-t-\pi+a) + pU(y-t-\pi+a-f) > U(y-t-\pi) \tag{3}$$

Similarly, an individual who would gain b would engage in it if[9]

$$(1-p)U(y-t-\pi+b) + pU(y-t-\pi+b-f) > U(y-t-\pi) \tag{4}$$

The proportion of the population engaging in the activity, n, is determined by (3) and (4).

Let EU_a and EU_b be, respectively, the expected utility of an individual, conditional on whether he gains a or b from engaging in the activity and where the decision whether to engage in it is made according to (3) or (4). Then the (unconditional) expected utility of an individual, calculated before he knows whether his potential gain is a or b, is

$$EU = qEU_a + (1-q)EU_b \tag{5}$$

The problem discussed here is the determination of the probability and fine that maximize expected utility.[10] Formally, the problem is to choose p and f to maximize (5), where (3) and (4) determine EU_a and EU_b in terms of p, f, t, and π, and where (1) and (2) determine π and t. In order to make this problem interesting, it is assumed that if it were costless to control individual behavior, then expected utility would be maximized if those who would gain the larger amount engage in the activity, but not those who would gain the smaller amount. In other words, $a < e < b$.[11]

Before proceeding it is necessary to discuss the *threshold probability*, the highest probability below which it is impossible to deter individuals from engaging in the activity Such a probability exists because there is a limit to how much an individual can be fined.[12] More precisely, let w be an individual's wealth, g the potential gain, and $\bar{p}(w, g)$ the threshold probability. Since g is nonmonetary in nature, if an individual engages in the activity and is fined the maximum amount w, his utility is $U(g)$. The threshold probability is determined by

$$(1-\bar{p})U(w+g) + \bar{p}U(g) = U(w) \tag{6}$$

since with the highest possible fine, $\bar{p}$ is such that the individual would be indifferent between engaging and not engaging in the activity. Obviously, if p is less than $\bar{p}$, the individual would engage in the activity since the fine cannot be raised.

[6]Subscripts are used to denote partial derivatives.

[7]The conclusions of this paper would not be affected if the cost also depended on the number of persons who engage in the activity.

[8]Per capita fine revenue is treated as deterministic; the justification is, of course, the law of large numbers.

[9]If individuals are indifferent between engaging and not engaging in the activity, it will be clear from context whether or not it is assumed that they engage in it.

[10]This problem is equivalent to that of maximizing the sum of utilities for a population with a proportion q of individuals who would gain a from engaging in the activity and a proportion $(1-q)$ who would gain b.

[11]If $e < a$, then there is no reason to discourage anyone from engaging in the activity. If $e > b$, then all should be discouraged if it is costless to do so; when it is costly to catch individuals, then the analysis can be easily developed from what follows.

[12]The existence of this probability was first noted by Michael Block and Robert Lind.

Solving for the threshold probability from (6) gives[13]

$$\bar{p}(w,g) = \frac{U(w+g) - U(w)}{U(w+g) - U(g)} \tag{7}$$

Note that the higher the private gain, the higher the threshold probability,

$$\bar{p}_g = [U'(w+g)[U(w) - U(g)] + U'(g)[U(w+g) - U(w)]] \div [U(w+g) - U(g)]^2 > 0 \tag{8}$$

This makes sense because it should be harder to prevent individuals from engaging in the activity if the gain from doing so rises.

II. Individuals are Risk Neutral

In this case, the optimal tradeoff between the probability and magnitude of fines is described by the following proposition.

PROPOSITION 1: *Suppose that individuals are risk neutral and that it is optimal to control the externality creating activity. Then the optimal probability is as low as possible (equal to the threshold probability of those who gain the least) and the optimal fine is as high as possible (equal to an individual's wealth).*[14]

(a) *This result is true no matter how low the cost of catching individuals.*

(b) *At the optimum, only those individuals whose private gains exceed the external cost engage in the activity.*

PROOF:

Let p^* and f^* be the optimal values of the probability and fine and suppose that $p^* > 0$ (i.e., it is optimal to control that activity). The proposition then states the $p^* = \bar{p}(y - t - \pi, a)$ and $f^* = y - t - \pi$.

Since $p^* > 0$, the *a*s (those who would gain a) do not engage in the activity and the *b*s (those who would gain b) do engage in it: otherwise there are three possibilities to consider. If all individuals engage in the activity, it would clearly be better to spend nothing to control the activity (i.e., set $p = 0$), contradicting the optimality of p^*.[15] If only the *a*s engage in the activity, it would again be better to spend nothing to control the activity since then the *b*s, for whom the private benefit exceeds the external cost, would also be induced to engage in the activity.[16] If neither the *a*s nor the *b*s engage in the activity, it is possible to alter only the fine so as to induce the *b*s but not the *a*s to engage in the activity.[17]

Suppose that $p^* < \bar{p}(y - t - \pi, a)$. Then by definition the *a*s would engage in the activity, contradicting the result of the previous paragraph.

Suppose that $p^* > \bar{p}(y - t - \pi, a)$. Since individuals are risk neutral, let $U(w) = w$ without loss of generality. Then, since the *a*s are not engaging in the activity and the *b*s are, it follows that

$$(1 - p^*)(y - t - \pi + a) + p^*(y - t - \pi + a - f^*) \leq y - t - \pi \tag{9}$$

$$(1 - p^*)(y - t - \pi + b) + p^*(y - t - \pi + b - f^*) \geq y - t - \pi \tag{10}$$

which, respectively, reduce to

[13] If $g \geq w$, define $\bar{p}(w,g) = 1$. In this case an individual would always choose to engage in the activity, since the gain is at least as high as the maximum potential loss through a fine. If $0 < g < w$, then $0 < \bar{p}(g, w) < 1$. In this case, there exists a probability and a fine which could deter individuals from engaging in the activity. It will be assumed that the latter case applies.

[14] It is clearly optimal to control the activity if the cost of catching individuals (at the threshold probability of those who gain the least) is sufficiently low.

[15] Since at p^* the *a*s and *b*s engage in the activity, $\pi = e$ and $t = c(p^*, \lambda) - p^*f$. Therefore, from (5), $EU = y - e + qa + (1 - q)b - c(p^*, \lambda)$ but at $p = 0$, $EU = y - e + qa + (1 - q)b$ which is larger since $c(p^*, \lambda) > 0$.

[16] Since only the *a*s engage in the activity, $\pi = qe$ and $t = c(p^*, \lambda) - p^*f^*q$. Therefore, from (5), $EU = y - qe - c(p^*, \lambda) + qa$. But at $p = 0$, $EU = y - e + qa + (1 - q)b$. The latter exceeds the former since $-e + (1 - q)b > -e + (1 - q)e = -qe > -qe - c(p^*, \lambda)$.

[17] If neither engage in the activity, $\pi = 0$ and $t = c(p^*, \lambda)$. Therefore from (5), $EU = y - c(p^*, \lambda)$. It is clear that there is an $f^0 < f^*$ such that only the *b*s engage in the activity. Then $\pi = (1 - q)e$ and $t = c(p^*, \lambda) - (1 - q)p^*f^0$. From (5), $EU = y - c(p^*, \lambda) + (1 - q)(b - e)$, which is higher than expected utility when neither engage since $b > e$.

$$(11) \qquad a \leq p^* f^*$$

$$(12) \qquad b \geq p^* f^*$$

There are two cases to consider. If $a < p^* f^*$, then holding f^* fixed, lower p^* slightly to p' so as to satisfy $a < p' f^*$. Then, since $b > p' f^*$, it is still true that (11) and (12) are satisfied at p' and f^*, that the as do not engage in the activity and that the bs do. Therefore at p' and f^*, π has not changed but t has changed to t', where

$$(13) \qquad t' = c(p', \lambda) - (1 - q) p' f^*$$

Next subtract expected utility (5) at p^* and f^* from that at p' and f^*, to get

$$\begin{aligned}(14) \quad & [q(y - t' - \pi) + (1 - q) \\ & (y - t' - \pi + b - p' f^*)] \\ & - [q(y - t - \pi) + (1 - q) \\ & (y - t - \pi + b - p^* f^*)] \\ & = (t - t') + (1 - q)(p^* - p') f^* \\ & = c(p, \lambda) - c(p', \lambda) > 0\end{aligned}$$

Thus, p^* and f^* could not have been optimal. The other case is $a = p^* f^*$. In this case,

$$\begin{aligned}(15) \qquad f^* &= a/p^* < a/\bar{p}(y - t - \pi, a) \\ &= y - t - \pi\end{aligned}$$

Thus, it is possible to raise f^* slightly to $f' < y - t - \pi$ and to lower p^* to $p' = p^* f^* / f'$. Since $p' f' = p^* f^*$, (11) and (12) continue to hold at p' and f', π does not change and taxes fall by

$$\begin{aligned}(16) \quad & [c(p^*, \lambda) - (1 - q) p^* f^*] \\ & - [c(p', \lambda) - (1 - q) p' f'] \\ & = c(p^*, \lambda) - c(p', \lambda) > 0\end{aligned}$$

Since taxes fall and nothing else has changed, expected utility must have risen, so that again p^* and f^* could not have been optimal. Hence $p^* = \bar{p}(y - t - \pi, a)$ as claimed and, by definition of the threshold probability, $f^* = y - t - \pi$; otherwise the as would engage in the activity.

III. Individuals are Risk Averse

In this case, the optimal tradeoff between the probability and magnitude of fines is described by two propositions. The first says that if the cost of catching individuals, λ, is sufficiently small, the optimal probability equals one. (Recall that if $\lambda = 0$, it is costless to catch individuals who engage in the activity and that as λ increases it becomes more costly to catch individuals for any p.) As noted in the introduction, this is because the disutility from bearing the risk of being caught is more important than the savings in the cost of catching individuals. When the optimal probability equals one, the optimal fine equals the private gain b of those who engage in the externality generating activity. (Although any fine between the external cost e and b would lead those individuals to engage in the activity, a fine of b is used in order to transfer income from those who have favorable opportunities for gains to those who do not.) The second proposition says that as λ increases, the optimal probability may never become as low as the threshold probability before it drops to zero and the optimal fine may never become as high as an individual's wealth. As noted in the introduction, this is because the use of a smaller probability and a higher fine may lower utility due to risk bearing and more than offset the benefits from controlling participation in the activity. These propositions are illustrated in Figure 1, which relates the optimal probability p^* to λ.[18] The diagram also shows the threshold probability, $\bar{p}(y - t - \pi, a)$, of individuals who would gain a from engaging in the activity.

PROPOSITION 2: *Suppose that individuals are risk averse. Then if the cost of catching them is sufficiently low, the optimal probability equals one and the optimal fine equals the private gain of those who engage in the externality creating activity.* (See the Appendix for the proof of Proposition 2.)

The proof consists of several steps. It is first shown that when $\lambda = 0$, $p^* = 1$. Then, in three

[18]The probability p^* does not necessarily decline with λ everywhere, although one would expect this to be the typical case. As λ rises, *ceteris paribus*, the tax necessary to maintain p at any level would rise, lowering wealth and raising absolute risk aversion (assuming it is decreasing with wealth). Since this would increase the disutility of bearing the risk of being caught, there would be a tendency for p^* to rise

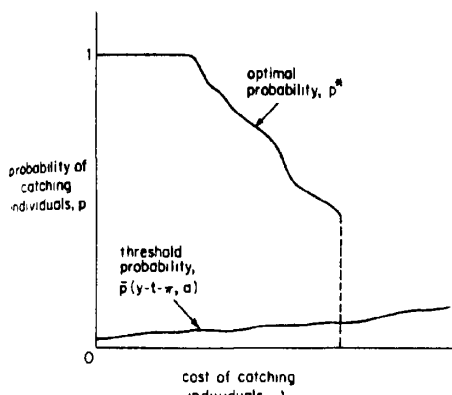



FIGURE 1

steps, a continuity argument is used to show that as $\lambda \to 0$, $p^* \to 1$. By a similar argument, it is shown that as $\lambda \to 0, f^* \to b$, where f^* is the optimal fine. Finally, these facts are used to prove the proposition.

PROPOSITION 3: *Suppose that individuals are risk averse. Then, as the potential gains from controlling the externality creating activity approach zero (i.e., as $b \to a$), the (minimum positive) optimal probability approaches one and the optimal fine approaches the private gain of those who engage in the activity.*

The reasoning in the proof is similar to that in Propositions 1 and 2 and will only be sketched below. Referring to Figure 2, the proof shows that given any small rectangle R (with northwest corner at (0, 1)), the positive portion of the $p^*(\lambda)$ schedule cannot be outside R if b is sufficiently close to a.

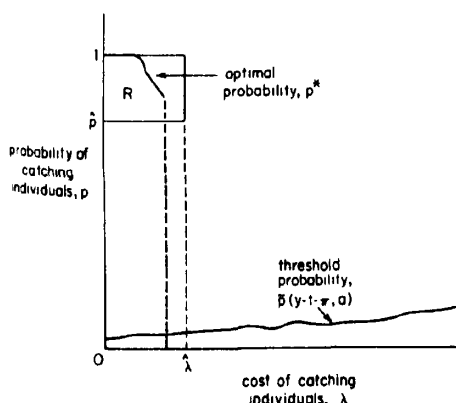



FIGURE 2

PROOF:

Let the coordinates of the southeast corner of R be $(\hat{\lambda}, \hat{p})$. If $p^*(\lambda) > 0$ is ever outside R, either $\lambda > \lambda$ or $p^*(\lambda) < \hat{p}$. Suppose first that $\lambda > \lambda$. Since $p^*(\lambda)$ is optimal, it must be at least equal to the threshold probability at λ, namely $\bar{p}(y - t - \pi, a)$, where t and π are implicitly determined by λ. Since this threshold probability is bounded away from zero,[19] expenditure on catching individuals is positive. Therefore EU is bounded away from $V(1, 0)$. (See proof of Proposition 2 for definition of V.) But clearly as $b \to a$, EU approaches $V(1, 0)$ if the activity is not controlled. Therefore, $p^*(\lambda)$ could not be optimal if b is sufficiently close to a.

Now assume that $p^*(\lambda) < \hat{p}$. Since $p^*(\lambda) > 0$ and is optimal, only the bs engage in the activity. But the fine is bounded from below, for otherwise the as would engage in the activity. Thus (refer to step (ii) of the proof to Proposition 2), disutility due to risk bearing is imposed on the bs; EU is therefore bounded away from $V(1, 0)$, which leads to a contradiction by the argument of the previous paragraph.

To show that the fine approaches b, note first that if the bs engage in the activity, the expected fine must be less than or equal to b, that is, the fine must be less than or equal to $b + b(1 - p)/p$. Since, as $b \to a$, the minimum positive optimal probability approaches 1, the fine is bounded above by $\lim_{p \to 1} b + b(1 - p)/p = b$. On the other hand, since the as do not engage in the activity, the fine must approach a limit at least equal to a, since the minimum positive probability approaches 1. Thus, as $b \to a$, the optimal fine approaches b (and a).

IV. Insurance Against Fines

Insurance against fines is in general not allowed in this country. However, because

[19]It is clear that the threshold probability for the as is bounded away from zero since income is bounded from above.

such insurance would reduce the disutility of bearing the risk of being fined for individuals who engage in the externality creating activity, it might seem that the insurance could be socially desirable under some conditions.[20] This section shows that such insurance cannot be beneficial because, as noted in the introduction, the government could achieve the same result by lowering the fine.

The type of insurance against fines to be considered is coverage which individuals purchase before they know their private gains from engaging in the activity. In other words, it is assumed that it would be impractical for individuals to buy insurance once they know their private gains.

Introduce the following notation:

h = insurance coverage against fines.

z = premium for coverage against fines.

It is assumed that insurance is sold either by competitive firms or directly by a public agency and, in either case, that the premium is actuarially fair. Thus, since n is the fraction of individuals who engage in the activity,

$$z = nph \tag{17}$$

To see that insurance cannot increase expected utility, consider an equilibrium in which individuals purchase coverage h against fines. If g is the gain from engaging in the activity, an individual would engage in it if

$$\begin{aligned}(1 - p)U(y - t - z - \pi + g) \\ + pU(y - t - z - \pi + g - f + h) \\ > U(y - t - z - \pi)\end{aligned} \tag{18}$$

Now consider the outcome if individuals are not allowed to purchase insurance against fines and the tax is changed from t to $\bar{t} = t + z$, the fine from f to $\bar{f} = f - h$, and nothing else is changed. Then an individual would engage in the activity if

$$\begin{aligned}(1 - p)U(y - \bar{t} - \pi + g) \\ + pU(y - \bar{t} - \pi + g - \bar{f}) \\ > U(y - \bar{t} - \pi)\end{aligned} \tag{19}$$

which is clearly equivalent to (18). Thus individual behavior and expected utility in this case are identical to that when individuals purchased coverage h. To verify that the budget balances, note that it must have been true in the equilibrium when h was purchased that $t = c - npf$. By this and (17),

$$\begin{aligned}\bar{t} &= t + z = c - npf + nph \\ &= c - np(f - h) = c - np\bar{f}\end{aligned} \tag{20}$$

which completes the argument.

V. Double Parking—A Numerical Example

The optimal tradeoff between the probability and magnitude of fines if individuals are risk averse may be illustrated by a numerical example. The example concerns the control of double parking in a hypothetical city in which the private gains to residents from double parking usually are exceeded by the congestion costs imposed, but occasionally are not. The example is described by the following data:

100,000 = population of the city

10,000 = number of locations where a resident could double park

25 = number of locations which each policeman can check per hour (and, if necessary, at which he must write a ticket)

s = wage of a policeman per hour

$\log w$ = utility of wealth w for each resident

\$10,000 = initial wealth of each resident

\$5.00 = congestion costs from double parking

\$4.50, \$50.00 = possible private gains to each resident from double parking

.90 = probability that the private gain is \$4.50.

In order to catch individuals who double park with probability p, $10{,}000p$ locations must be checked, $10{,}000p/25 = 400p$ policemen must be employed, and a total of $400ps$ must be spent on enforcement. If the fine is f,

[20]Closely related questions which arise in connection with liability insurance are discussed by Shavell.

TABLE 1—THE CONTROL OF DOUBLE PARKING

Wage (Per Hour)	Optimal Probability	Optimal Fine	Total Enforcement Expenditures at Optimal Probability	Threshold Probability	Total Enforcement Expenditures at Threshold Probability
$3.00	.09	$50.10	$108	.00006	$0.07
$5.00	.07	$64.10	$140	.00006	$0.12
$7.50	.06	$74.80	$180	.00006	$0.18
$10.00	.05	$89.80	$200	.00006	$0.23
$15.00	04	$111 90	$240	.00006	$0.35
$25.00	.03	$149.00	$300	.00006	$0 58

then total fine revenue is pf times the number of residents who double park. Per capita taxes equal total expenditures on enforcement minus total fine revenue, all divided by 100,000. Per capita congestion costs are $5.00 times the number of people who double park, divided by 100,000. Given his wealth net of taxes and congestion costs, his private gain, and the probability and magnitude of the fine, each resident decides whether to double park (see Section I). The expected utility of a typical resident may then be calculated.

The probability and fine which maximize expected utility of the residents were computed for various values of the policemen's wage. The results may be summarized as follows:

(a) If the wage is less than $2.83, the optimal probability is 1 and the optimal fine is $50. This result, which illustrates Proposition 2, shows that when policemen can be cheaply hired, it is best to employ a sufficient number to catch all individuals who double park. Total expenditures on enforcement may be as high as $1,132 (when the wage is $2.83). In contrast, the government could use the maximum possible fine of $10,000 with the threshold probability of .00006. This is the lowest probability which would deter those who would gain $4.50 from double parking; however, those who would gain $50 would double park. Although this system of enforcement would involve total expenditures of at most $0.07, it is not optimal because of the high risk imposed on those who double park.

(b) As the wage rises above $2.83, the optimal probability rapidly declines and the fine increases. Table 1 reports the results for several reasonable values of a policeman's wage. As the wage increases from $3 to $25, the optimal probability falls from .09 to .03, the optimal fine rises from $50.10 to $149.00, and total enforcement expenditures increase from $108 to $300. In contrast, if the fine were incorrectly set at its maximum $10,000, and the probability were set at the threshold .00006, then total expenditures on enforcement would never exceed $0.58.

(c) As the wage increases to extremely high levels, the optimal probability declines to a value slightly above the threshold probability and the optimal fine rises to approximately $9,000. Only when the wage exceeds $200,000 *per hour* is the optimal policy to do nothing about double parking. This illustrates Proposition 3 since, when it is optimal to control double parking, the optimal fine never reaches $10,000 and the optimal probability never becomes as low as the threshold probability. If the parameters of the example were different, the wage at which it first becomes optimal to do nothing about double parking might be much lower; as a result, the maximum optimal fine might be much lower and the minimum positive optimal probability much higher. For instance, this would be the case if the lower private gain from double parking were just below the congestion costs imposed.

VI. Concluding Remarks

The model used here abstracted from a variety of considerations relevant to the deter-

mination of the optimal probability and fine. Several of these are now mentioned.

It was assumed that the private gains from engaging in the externality creating activity could not be observed. However, in some contexts the private gains might be identifiable at little cost. If this is the case, the fine (and possibly the probability) could depend on the private gain. Those who would gain less than the external cost could be discouraged from engaging in the activity by setting their fine sufficiently high, and those who would gain more than the external cost could be induced to engage in the activity by setting their fine sufficiently low, possibly at zero. Therefore, the optimal probability can be lowered since disutility due to risk bearing can be reduced.

It was also assumed in the model that individuals who did not engage in the activity were never mistakenly fined. If this possibility had been taken into account, the conclusions of this paper would be reinforced, since all individuals would then bear the risk of paying a fine.

Furthermore, it was assumed that if individuals engaged in the activity at all, they did so at a particular level. A more general model would allow individuals to engage in the activity at varying levels. Similarly, it was assumed that the distribution of private gains was discrete, whereas in general the distribution might be continuous. Neither of these extensions would affect the basic results of this paper. All that is required for the results is that, given the optimal probability and fine, some risk-averse individuals generate externalities at a positive level and are subject to the risk of having to pay a fine.

Finally, it was assumed that individuals had the same level of wealth. However, differences in wealth may be important in many situations. Suppose that absolute risk aversion decreases with wealth and that the probability and fine cannot be made to depend on wealth. Then, *ceteris paribus*, any given probability and fine would be less likely to discourage a wealthy individual from engaging in the activity than a poor one. As a result, in the optimal solution, some wealthy individuals might be underdeterred—induced to engage in the activity even though their private gains are less than the external cost—and some lower income individuals who are able to pay the fine might be overdeterred. However, some poor individuals who are unable to pay the fine might be underdeterred.

APPENDIX

PROOF of Proposition 2:

Let p^* and f^* be the optimal values of the policy parameters. Define $V(p, \lambda)$ as the maximum expected utility (5) given p and λ.

(i) $V(1, 0) > V(p, 0)$ for $p < 1$: If, given p and the associated optimal fine, f, all individuals engage in the activity, then $\pi = e$ and $t = -pf$, so that

$$\text{(A1)}\quad \begin{aligned} EU_a &= (1 - p)U(y + pf - e + a) \\ &\quad + pU(y + pf - e + a - f) \\ &\leq U(y + pf - e + a - pf) \\ &= U(y - e + a) \end{aligned}$$

The inequality follows since U is concave (individuals are risk averse). Similarly,

$$\text{(A2)}\quad EU_b \leq U(y - e + b)$$

Hence

$$\text{(A3)}\quad \begin{aligned} V(p, 0) &\leq qU(y - e + a) \\ &\quad + (1 - q)U(y - e + b) \\ &\leq U(y - e + qa \\ &\qquad + (1 - q)b) \\ &= U(y + (1 - q)(b - e) \\ &\qquad + q(a - e)) \end{aligned}$$

On the other hand, if the probability equals 1 and the fine equals b, then the as will not engage in the activity and the bs will. Hence $\pi = (1 - q)e$, $t = -(1 - q)b$, and

$$\text{(A4)}\quad \begin{aligned} EU_a &= EU_b \\ &= U(y + (1 - q)(b - e)) \end{aligned}$$

so that, since $a < e$,

$$\text{(A5)}\quad \begin{aligned} V(1, 0) &\geq U(y + (1 - q)(b - e)) \\ &> U(y + (1 - q)(b - e) \\ &\qquad + q(a - e)) \geq V(p, 0) \end{aligned}$$

If at p and f no individuals engage in the activity, then $\pi = t = 0$, so

$$\text{(A6)}\quad V(p,0) = U(y) < U(y + (1-q)(b-e)) \leq V(1,0)$$

If at p and f, only the as engage in the activity, then $\pi = qe$, $t = -qpf$; the following chain of inequalities holds:

$$\text{(A7)}\quad V(p,0) = q[(1-p)U(y + q(pf-e) + a) + pU(y + q(pf-e) + a - f)] + (1-q)U(y + q(pf-e)) \leq qU(y + q(pf-e) + a - pf) + (1-q)U(y + q(pf-e)) \leq U(y + q(pf-e) + q(a-pf)) = U(y + q(a-e)) < U(y + (1-q)(b-e)) \leq V(1,0)$$

The first two inequalities follow from the concavity of U; the third inequality follows from $a < e < b$; and the last inequality follows from (A5).

The remaining possibility is that at p and f the as do not engage in the activity and the bs do engage in the activity. Then $\pi = (1-q)e$ and $t = -(1-q)pf$. In this case, raise the probability to 1 and lower the fine to a value f^0 such that the utility of the bs is unchanged if π and t are held constant; i.e.,

$$\text{(A8)}\quad (1-p)U(y - t - \pi + b) + pU(y - t - \pi + b - f) = U(y - t - \pi + b - f^0)$$

At a probability equal to 1 and a fine of f^0, the fine revenue raised by the government is higher since

$$\text{(A9)}\quad f^0 > pf$$

To see this, observe that the left-hand side of (A8) is less than $U(y - t - \pi + b - pf)$ since U is concave (and $f \geq a > 0$, for otherwise the as would engage in the activity), so that (A8) cannot hold unless the fine f^0 exceeds pf. It is also true that

$$\text{(A10)}\quad f^0 \leq b$$

since otherwise the left-hand side of (A8) is less than $U(y - t - \pi)$, contradicting the assumption that the bs choose to engage in the activity.

Now there are two cases to consider: $f^0 \geq a$ and $f^0 < a$. In the former case, let per capita taxes be reduced by the amount

$$\text{(A11)}\quad s_1 = (1-q)(f^0 - pf) > 0$$

which is positive by (A9). At a probability equal to 1, f^0, $\pi = (1-q)e$, and the new and lower taxes, it is still true that the as do not engage in the activity and the bs do, for $a \leq f^0 \leq b$. However, the as are better off than at $p < 1$ for after-tax income has risen by s_1; and the bs are better off because $U(y - t - \pi + b - f^0)$ is less than $U(y - t + s_1 - \pi + b - f^0)$. Consequently p and f could not have been optimal.

On the other hand, if $f^0 < a$, then at a probability equal to 1, the bs will still choose to engage in the activity; although the as would in fact also choose to engage in the activity, note that if they did *not*, their expected utility at a probability of 1 and f^0 (and $\pi = (1-q)e$, $t = -(1-q)pf$) would be identical to that at p and f. Now reduce taxes by s_1, as in the previous case. This raises both the as and bs expected utility assuming (temporarily) that the as do not engage in the activity. Expected utility, EU, is therefore

$$\text{(A12)}\quad qU(y - t + s_1 - \pi) + (1-q) \cdot U(y - t + s_1 - \pi + b - f^0)$$

Now at a probability of 1, raise the fine from f^0 to a. At this level, the as do not engage in the activity and the bs do, so that $\pi = (1-q)e$. Since per capita fine revenue rises from $(1-q)f^0$ to $(1-q)a$, per capita taxes can be reduced further by

$$\text{(A13)}\quad s_2 = (1-q)(a - f^0) > 0$$

Define $s = s_1 + s_2$. Expected utility is therefore

$$\text{(A14)}\quad qU(y - t + s - \pi) + (1-q) \cdot U(y - t + s - \pi + b - a)$$

If (A14) exceeds (A12), it will have been demonstrated that EU can be made higher at

a probability of 1 than at p, completing this step of the proof. But expected wealth in (A12) and (A14) is easily verified to be identical. Hence, since U is concave, and

$$\text{(A15)}\quad (y - t + s - \pi) > (y - t + s_1 - \pi)$$
$$(y - t + s - \pi + b - a) < (y - t + s_1 - \pi + b - f^0)$$

(A14) exceeds (A12).[21]

(ii) For any $\delta > 0$ there exists a $\gamma > 0$ such that if $p < 1 - \delta$, then $V(1,0) - V(p,0) > \gamma$: This step follows directly from the previous step. Recall that at $p < 1$ (and in particular at $p < 1 - \delta$) there were four possibilities. If all individuals engage in the activity, it was shown that $V(1,0) > V(p,0)$ by at least a positive amount, say γ_1, which is independent of p (see (A5)). If no individuals engage in the activity, or only the as engage in it, again $V(1,0) > V(p,0)$ by at least positive amounts γ_2 and γ_3, respectively, independent of p (see (A6) and (A7), respectively). Finally, if at p the as do not engage in the activity and the bs do, then it was shown that $V(1,0) > V(p,0)$. The first part of that demonstration involves raising the probability to one and lowering the fine to f^0; this generated surplus s_1 (see (A11)), which was distributed and raised expected utility. That argument implies that the size of s_1—and hence of the subsequent utility gain—is larger the smaller is p. Since p is no larger than $1 - \delta$, there exists a $\gamma_4 > 0$ serving as a lower bound for $V(1,0) - V(p,0)$ in this final case. Pick γ = minimum γ_i.

(iii) For any $\delta > 0$, there exists $\epsilon > 0$ such that if $\lambda < \epsilon$, the optimal probability, $p^*(\lambda)$, satisfies $p^*(\lambda) \geqq 1 - \delta$; in other words, $p^*(\lambda) \rightarrow 1$ as $\lambda \rightarrow 0$: Assume otherwise. Then for some $\delta > 0$, there exists a λ arbitrarily small such that $p^*(\lambda) < 1 - \delta$. It is easy to show that $V(1,\lambda) \rightarrow V(1,0)$ as $\lambda \rightarrow$ (Therefore, pick ϵ such that if $\lambda < \epsilon$, the $|V(1,\lambda) - V(1,0)| < \gamma/2$. Now certainl $V(p,\lambda) \leqq V(p,0)$ for all p and since for p . $1 - \delta$, $V(1,0) - V(p,0) > \gamma$ (by step (ii)) i must be true that $V(1,0) - V(p,\lambda) >$ ⁊ Hence, if $\lambda < \epsilon$, $V(1,\lambda) - V(p,\lambda) > \gamma/2 >$ (Therefore $V(p,\lambda)$ could not have been th maximum, a contradiction.

(iv) If $p^*(\lambda) = 1$, the optimal fine is b: I $p = 1$ and $f = b$, the bs engage in the activit the as do not, and the utility of each indivic ual is $U(y - c(1,\lambda) + (1 - q)(b - e))$. Thi fine must be optimal since it coincides wit the first best solution to the problem.[22]

(v) $p^*(\lambda) = 1$ for all λ sufficiently lov It is first shown that $V_p(1,0) > 0$. To do thı consider the function $W(p) = \max_f Y(p,f$ where

$$\text{(A16)}\quad Y(p,f) = qU(y + (1 - q)(pf - e) + (1 - q)[(1 - p)U(y + (1 - q)(pf - e) + b) + pU(y + (1 - q)(pf - e) + b - f)$$

Thus $W(p)$ is the expected utility of indivıdı als as a function of p when $\lambda = 0$ and f i selected optimally, but without constraining to be such that the bs are induced to engage i the activity and the as discouraged fror doing so. Therefore,

$$\text{(A17)}\qquad W(p) \geqq V(p,0)$$

It may easily be verified that

$$\text{(A18)}\quad W(1) = Y(1,b) = V(1,0)$$

Thus, if $W'(1) > 0$, then $V_p(1,0) > 0$. Bu since $W(1) = Y(1,b)$ and $Y_f(1,b) = 0$ (this i the first-order condition for optimal selectioı of f), it follows that

[21]In general, if $A = qU(w_1) + (1 - q)U(w_2)$ and $B = qU(w_3) + (1 - q)U(w_4)$, where $qw_1 + (1 - q)w_2 = qw_3 + (1 - q)w_4$, $w_1 < w_2$, $w_3 < w_4$, $w_3 > w_1$, and $w_4 < w_2$, then $A < B$ if U is concave. To prove this, define $C(t) = qU(w_1 + t) + (1 - q)U(w_2 - qt/(1 - q))$. Then $A = C(0)$ and $B = C(t_0)$ for some $t_0 > 0$. Now $C'(t) = q[U'(w_1 + t) - U'(w_2 - qt/(1 - q))] > 0$ as long as $w_1 + t < w_2 - qt/(1 - q)$ by concavity of U. Hence $B > A$ given our assumptions.

[22]Consider the first best (benevolent dictator's) prot lem of maximizing expected utility given that p = subject only to the constraint that resources balance, y - $ne - c(1,\lambda) = qy_a + (1 - q)y_b$, where y_a is the wealth c the as and y_b that of the bs. It is clear that the solutio involves ordering only the bs to engage in the activity an selecting y_a and y_b so that the marginal utilities of wealt for the as and bs are equal. Thus $n = (1 - q)$, $y_a = y$ - $c(1,\lambda) + (1 - q)(b - e)$ and $y_b = y_a - b$ so that EU_b = $U(y_a - b + b) = U(y_a) = EU_a$.

$$\text{(A19)} \quad W'(1) = Y_p(1,b) + Y_f(1,b)\frac{df}{dp} = Y_p(1,b)$$
$$= q(1-q)bU'(y+(1-q)(b-e)) + (1-q)[U(y+(1-q)(b-e)) - U(y+(1-q)(b-e)+b) + (1-q)bU'(y+(1-q)(b-e))] > 0$$

The inequality follows because, by concavity of U, $U(y+(1-q)(b-e)) - U(y+(1-q)(b-e)+b) > -bU'(y+(1-q)(b-e))$.

Suppose it is not true that $p^*(\lambda) = 1$ for all λ sufficiently small. Then, since it was shown in step (iii) that $p^*(\lambda) \to 1$ as $\lambda \to 0$, there must exist a sequence $\{\lambda_i\}_{i=1}^{\infty}$ where $\lambda_i \to 0$, $p^*(\lambda_i) < 1$, and $p^*(\lambda_i) \to 1$. Since $p^*(\lambda_i) < 1$, it is an interior optimum so that $V_p(p^*(\lambda_i), \lambda_i) = 0$ for all i. Therefore,

$$\text{(A20)} \quad \lim_{i \to \infty} V_p(p^*(\lambda_i), \lambda_i) = 0$$

On the other hand, by continuity of $V_p(p, \lambda)$,

$$\text{(A21)} \quad \lim_{i \to \infty} V_p(p^*(\lambda_i), \lambda_i) = V_p(1,0) > 0$$

which is a contradiction.

REFERENCES

G. S. Becker, "Crime and Punishment: An Economic Approach," *J. Polit. Econ.*, Mar./Apr. 1968, *76*, 169–217.

M. K. Block and R. C. Lind, "Crime and Punishment Reconsidered," *J. Legal Stud.*, Jan. 1975, *4*, 241–47.

Kenneth G. Elzinga and William Breit, *The Antitrust Penalties: A Study in Law and Economics*, New Haven 1976.

H. J. Goldschmid, "An Evaluation of the Present and Potential Use of Civil Money Penalties as a Sanction by Federal Administrative Agencies," in *Recommendations and Reports of the Administrative Conference of the United States: July 1, 1970–December 31, 1972*, Vol. 2, Washington 1973.

S.-C. Kolm, "A Note on Optimum Tax Evasion," *J. Publ. Econ.*, July 1973, *2*, 265–70.

W. M. Landes and R. A. Posner, "The Private Enforcement of Law," *J. Legal Stud.*, Jan. 1975, *4*, 1–46.

Richard A. Posner, *Economic Analysis of Law*, 2d ed., Boston 1977.

G. C. Rausser and R. Howitt, "Stochastic Control of Environmental Externalities," *Annals Econ. Soc. Measure.*, Spring 1975, *4*, 271–92.

A. Samuels, "The Fine: The Principles," *Criminal Law Rev.*, 1970, 201–10; 268–72.

S. Shavell, "Accidents, Liability, and Insurance," disc. paper no. 685, Harvard Inst. Econ. Res., Feb. 1979.

B. Singh, "Making Honesty the Best Policy," *J. Publ. Econ.*, July 1973, *2*, 257–63.

T. N. Srinivasan, "Tax Evasion: A Model," *J. Publ. Econ.*, Nov. 1973, *2*, 339–46.

G. J. Stigler, "The Optimum Enforcement of Laws," *J. Polit. Econ.*, May/June 1970, *78*, 526–36.

Gordon Tullock, *The Logic of the Law*, New York 1971.

L. Weiss, "The Desirability of Cheating Incentives and Randomness in the Optimal Income Tax," *J. Polit. Econ.*, Dec. 1976, *84*, 1343–52.

Translog Flexible Functional Forms and Associated Demand Systems

By PETER SIMMONS AND DANIEL WEISERBS*

Laurits Christensen, Dale Jorgenson, and Lawrence Lau (hereafter C-J-L) have introduced translog direct and indirect preferences and the associated systems of demand functions (see also Jorgenson and Lau). These preferences are of interest for their own sake but are especially significant because of their properties as a flexible functional form. It is argued that they can represent arbitrary well-behaved preferences in the neighborhood of a given point with an accuracy of the second order. This suggests, as Lau has claimed in his 1974 paper, that concern directed towards precise functional specification may be misplaced since a flexible functional form can always be used. Moreover C-J-L proposed a new methodology for testing the implications of demand theory. They calculate the restrictions on the approximating translog function corresponding to the restrictions that demand theory imposes on unknown true preferences at any given point of approximation. They then proceeded to test these restrictions on the approximating translog demand functions

The purpose of this paper is to examine this proposed methodology and the properties of the indirect translog system. The main results are: (a) With a utility approximation it is not theoretically possible to discriminate between the hypothesis that translog preferences hold globally and the hypothesis that the true (but unspecified) preferences satisfy integrability conditions at the base point of approximation. (b) It is possible to define equally accurate approximations at the level of the demand equations or at the level of the marginal rates of substitution which do permit this distinction and which involve different parameter restrictions within the same estimating form. It is thus not possible to test the restrictions of demand theory on true preferences at a given point independently of the method of approximation selected. (c) These alternative approximations also have the advantage of allowing distinct tests of homogeneity and symmetry. (d) Empirically, homogeneity of degree zero of the budget shares is a more questionable hypothesis than integrability of either the true demand system or the approximate demand system. (e) The proposed methodology has important econometric implications.

I. Concepts and Methodology

Let $u = u(x_1, \ldots, x_n)$ be a twice continuously differentiable utility function representing unknown true preferences. Denote by q_i the quantity of commodity i, by p_i its price, and by y total expenditure (income) with the assumption that $q_i, p_i, y > 0$ for all i. Define x_i either equal to q_i or to the ratio p_i/y. In the first case, $u(x)$ is interpreted as the direct utility function; in the second case as the indirect utility function.

Given $u(x)$, define $z_i = f_i(x_i)$ and $\psi(z)$ by the identity $\psi(z) = u(x)$ for all x. Using the superscript A to denote an approximating function, let $\psi^A(z)$ be a second-order Taylor approximation of $\psi(z)$ at a base point $\bar{z}$:

$$(1)\quad \psi^A(z) = \alpha_o + \sum_i \alpha_i z_i + \frac{1}{2}\sum_i \sum_j \beta_{ij} z_i z_j$$

Since $\psi(z)$ is twice continuously differentiable by assumption, we have

$$(2a)\quad \alpha_o = \psi(\bar{z}) - \sum_i \frac{\partial\bar{\psi}}{\partial z_i}\bar{z}_i + \frac{1}{2}\sum_i \sum_j \frac{\partial^2\bar{\psi}}{\partial z_i \partial z_j}\bar{z}_i\bar{z}_j$$

*Catholic University of Louvain. We are grateful to the managing editor and to an anonymous referee for their comments on an earlier draft. Thanks are also due to M. Desai, E. de Souza, J.H. Dreze, and A. Magnus for helpful discussions and to D. Van Grunderbeeck for computational assistance. It is unlikely that they will accept responsibility for remaining errors.

(2b) $$\alpha_i = \frac{\partial \bar{\psi}}{\partial z_i} - \sum_j \frac{\partial^2 \bar{\psi}}{\partial z_i \partial z_j} \bar{z}_j$$

(2c) $$\beta_{ij} = \frac{\partial^2 \bar{\psi}}{\partial z_i \partial z_j}$$

Now, define the budget share functions corresponding to $u(x)$ (or $\psi(z)$) by $w_i = g_i(p, y)$ and the budget share functions corresponding to $u^A(x)$ (or $\psi^A(z)$) by $w_i = g_i^A(p, y; \alpha, \beta)$.

Then g_i satisfies the following so-called general restrictions where η_i and η_{ij} are income and price elasticities, respectively,

(3a) $\sum_i w_i = 1$ Adding up

(3b) $\sum_i w_i \eta_i = 1$ Engel aggregation

(3c) $\sum_i w_i \eta_{ij} = -w_j$ Cournot aggregation

(3d) $\sum_j \eta_{ij} = -\eta_i$ Homogeneity

(3e) $w_i(\eta_{ij} + w_j \eta_i) = w_j(\eta_{ji} + w_i \eta_j)$ Symmetry $i \neq j$

and negative semidefiniteness of the matrix of the substitution effects. Notice that these restrictions are not independent. For instance (3a), (3d), and (3e) imply (3b) and (3c).

The philosophy underlying this approach is that consumer plans are exactly determined by the maximization of $u(x)$ subject to the budget constraint and that $u(x)$ can be approximated by $u^A(x)$ to the second order. It follows from the properties of a Taylor approximation that $\psi^A(\bar{z}) = \psi(\bar{z})$ and that $g_i^A(\bar{z}) = g_i(\bar{z})$. In principle, (2) is only valid at $\bar{z}$. More attention will be devoted to the error components in the econometric section.

The methodology proposed by C-J-L for testing the validity of utility maximization applied to u at $\bar{z}$ is to estimate g_i^A, then use (2) to express the restrictions (3) on g_i at $\bar{z}$ in terms of the parameters of g_i^A. In this case, where the true preferences, but not necessarily the approximating preferences, satisfy a restriction at $\bar{z}$, C-J-L use the terminology "restriction" (hereafter called true restriction).

On the other hand, one can consider testing utility maximization applied to u^A by estimating g_i^A and imposing the restrictions based on (3). In this case C-J-L use the terminology "explicit restriction." The assumption being tested here is that preferences are translog and u^A represents the correct specification of consumer preferences.

Given that the approximation is exact at one point, it is only possible to test the true restrictions on u at that point. But explicit restrictions on u^A can be tested over any subset of the data desired and will be considered globally, that is, at all points of the sample.

One should realize the the distinction between testing true and explicit preferences is of great importance. The standard procedure in applied demand analysis has been to specify a function (utility, expenditure, or budget share) assumed to represent consumer behavior globally. In such a framework one can test general or particular restrictions[1] or compare the performances of various specifications although this is less frequently done. If a restriction or a specification is rejected still no conclusion can be reached concerning the validity of the theory. One could always hope that some other specification would perform better.

In C-J-L's approach z_i is set equal to the logarithm of x_i, thus $z_i = ln(p_i/y)$ for the indirect utility function and $z_i = ln\, q_i$ for the direct utility function. The corresponding budget share functions are

(4) $$w_i = \frac{\alpha_i + \sum_j \beta_{ij} z_j}{\alpha_m + \sum_j \beta_{mj} z_j}$$

where $\alpha_m = \Sigma_i \alpha_i$ and $\beta_{mj} = \Sigma_i \beta_{ij}$

Now if u exists and is twice continuously differentiable as assumed, its Hessian is symmetric at $\bar{z}$, i.e., $\beta_{ij} = \beta_{ji}$ in both the direct and the indirect cases. C-J-L then proceed by estimating the system (4) with β_{ij} free[2] and

[1]See for instance the surveys of Anton Barten, and Alan Brown and Angus Deaton.

[2]In the unconstrained case, C-J-L did not impose equality: the β_{ij} appearing in the various budget share equations were not constrained to have the same value.

with the matrix $[\beta_{ij}]$ constrained to be symmetric. On the basis of a likelihood-ratio test, they reject symmetry of the Hessian and consequently the validity of the theory of demand.

Thus C-J-L's methodology appears to allow testing of consumer theory independently of the specification of a functional form for true preferences. However their application raises the following questions:

1. With a second-order Taylor approximation to the utility function, how can we distinguish between symmetry of the true and the approximate Hessian?

2. Is symmetry of the Hessian the proper test of integrability of the demand functions?

3. Is the utility function the proper function to approximate in order to test the validity of the theory of demand?

4. If not, what would be a more adequate procedure?

5. What are the econometric implications of the distinction between g_i and g_i^A and consequently how should one estimate the latter?

6. In conclusion, what is the pertinence of C-J-L's results?

II. On Testing the Symmetry of the Hessian

The symmetry test is crucial in determining the validity of the theory of demand.[3] However, it is clear from (2c) that the Hessian of the translog utility function is constant independently of the data. Therefore, by the properties of a Taylor-series aproximation at a base point $\bar{z}$, one has $u_{ij}(\bar{z}) = u_{ij}^A(\bar{z})$ and consequently $u_{ij}(\bar{z}) = u_{ij}^A(z)$ for any z. It then follows that acceptance or rejection of symmetry of the true Hessian at one point implies acceptance or rejection of symmetry of the translog Hessian at all points. In other words, in this framework, when $\beta_{ij} = \beta_{ji}$ it is not possible to distinguish between the following two cases:

1) the translog utility function is an approximation to an unknown true function that has a symmetric Hessian at a base point;

2) the translog utility function represents consumer preferences globally.

Thus the answer to question 1 would be.

Using a Taylor approximation to the utility function prevents distinct tests of symmetry of the true and the approximate Hessian.[4]

Nevertheless, C-J-L concentrate on testing symmetry of the Hessian rather than of the substitution effects. However, we know that true consumer preferences exist (under certain regularity conditions) iff the true matrix of substitution effects is symmetric and negative semidefinite. So if the latter holds but $\beta_{ij} = \beta_{ji}$ is rejected, this merely means that consumer's true preferences cannot be represented globally by a translog utility function. It does not mean that well-behaved consumer preferences do not exist nor even that these preferences do not generate the budget share system (4). Indeed, there exist other utility functions that are not monotonic transformations of u^A, generating budget share systems of the same functional form as g_i^A but involving a different symmetry restriction on the Hessian in terms of the parameters of g_i^A.

The only utility functions different from the translog (1) yielding the budget share system (4) are:[5]

$$(5a) \quad u = \Big(a_o + \sum_j a_j z_j\Big)^{c_1} \Big(b_o + \sum_j b_j z_j\Big)^{c_2}$$

with $c_1 \neq c_2 \neq 0$ and $\alpha_i = c_1 a_i b_o + c_2 a_o b_i$, $\beta_{ij} = c_1 a_i b_j + c_2 a_j b_i$;

[3]Of course it is not the only test. Second-order conditions (negativity), unfortunately difficult to impose a priori in the translog context, are ignored both by C-J-L and ourselves.

[4]Probably because C-J-L did not realize this, their distinction between true and explicit additivity is erroneous. In fact their concept of true additivity is the ordinal definition $\partial(w_i/w_j)/\partial z_n = 0$ while their concept of explicit additivity is the cardinal one ($\psi_{ij} = \beta_{ij} = 0$) A similar confusion is made for homotheticity

[5]Incidentally (5a) can be viewed as a special case of Chisholm's generalization of Padé approximants to $\psi(z)$ when c_2/c_1 is a negative integer. We are indebted to Magnus for pointing this out and for providing a formal demonstration that (1) and (5) constitute the exhaustive set of utility functions yielding (4).

(5b) $$u = c \ln (a_o + \sum_j a_j z_j) + \frac{b_o + \sum_j b_j z_j}{a_o + \sum_j a_j z_j}$$

with $\alpha_i = a_o b_i + a_i(ca_o - b_o)$; $\beta_{ij} = a_j b_i + a_i(ca_j - b_j)$;

(5c) $$u = c \ln (a_o + \sum_j a_j z_j) + b_o + \sum_j b_j z_j$$

with $\alpha_i = ca_i + a_o b_i$; $\beta_{ij} = a_j b_i$.

Of course each of these utility functions has a symmetric Hessian but each also has $\beta_{ij} \neq \beta_{ji}$ Thus the answer to question 2 would be:

Testing the symmetry of the approximate Hessian is not obviously a proper test of the hypothesis of integrability of the true demand function.

Indeed, the budget share system (4) exhibits symmetry of the substitution effect iff:

(6) $$\beta_{ij} - w_i \beta_{mj} - w_j \beta_{im} = \beta_{ji} - w_j \beta_{mi} - w_i \beta_{jm}$$

In principle, one can test whether true symmetry is satisfied at an arbitrary base point using the relationships between the β_{ij} and the derivatives of true preferences at that point. Similarly one can test explicit symmetry using system (6) directly. Unfortunately with the translog utility approximation both tests coincide and (6) reduced to $\beta_{ij} = \beta_{ji}$.

III. On Approximating Ordinal Functions

One readily sees from equations (3) and (4) that homogeneity is incorporated in the model. This point is obvious. In the direct case, prices and income do not appear among the independent variables. In the indirect case, use is made of Roy's identity which implies homogeneity of degree zero of the budget share. Therefore the answer to question 3 would be:

a) In the C-J-L procedure one cannot conclude that rejection of symmetry is due solely to symmetry and not to homogeneity.

On the other hand the translog approximation and the relationship of the parameters to the utility function given by (2) are cardinal. However the restrictions of the theory are ordinal. Nevertheless, it can easily be shown that

b) The restrictions on true demands yield identical restrictions on the parameters for any increasing monotonic transformation of the utility function.

We shall now examine approximation of ordinal functions. Major reasons for doing so are: to admit both translog and nontranslog preferences by concentrating on the integrability conditions (6); to distinguish true and explicit restrictions; and to permit distinct tests on homogeneity and symmetry.

A. *Approximation of the Expenditure Equations*

Economists are accustomed to think in terms of expenditures which therefore seem the natural dependent variables to select. Let $p_i q_i = f_i(p, y)$ be the true expenditure equation assumed to be continuously differentiable and to satisfy $\Sigma_i f_i(p, y) = y$. Next define $h_i(\ln p, \ln y) = f_i(p, y)$ and approximate $h_i(\cdot)$ to the first order around $(\bar{p}, \bar{y})$ by

(7) $$h_i^A(\ln p, \ln y) = \alpha_i + \sum_j \beta_{ij} \ln p_j + \gamma_i \ln y$$

so that

(8) $$\beta_{ij} = \frac{\partial \overline{p_i q_i}}{\partial \ln p_j}; \quad \gamma_i = \frac{\partial \overline{p_i q_i}}{\partial \ln y};$$

$$\alpha_i = \overline{p_i q_i} - \sum_j \beta_{ij} \ln \bar{p}_j - \gamma_i \ln \bar{y}$$

The approximate budget share system is then

(9) $$w_i = \frac{a_i + \sum_i \beta_{ij} \ln p_j + \gamma_i \ln y}{\alpha_m + \sum_j \beta_{mj} \ln p_j + \gamma_m \ln y}$$

which can be viewed as a direct first-order approximation to the true budget shares

$$w_i = \frac{f_i(p, y)}{\sum_i f_i(p, y)}$$

Restrictions on true preferences at $(\bar{p}, \bar{y})$ are then derived using (8). At $(\bar{p}, \bar{y})$ the true functions have to satisfy the following conditions:

$$\gamma_i = \bar{y}\bar{w}_i - \beta_{im}$$

(homogeneity of degree one)

$$\alpha_m = \bar{y}(1 - ln\,\bar{y}) \qquad \text{(adding up)}$$

$$\beta_{ij} - \bar{w}_j\beta_{im} = \beta_{ji} - \bar{w}_i\beta_{jm}$$

(symmetry of the substitution effects)

plus the condition of negative semidefinitiveness of the matrix $[(p_ip_j)^{-1}(\beta_{ij} + \bar{y}\bar{w}_i\bar{w}_j - \beta_{im}\bar{w}_j)]$. It follows that $\beta_{mj} = 0$ (Cournot aggregation) and $\gamma_m = \bar{y}$(Engel aggregation).

On the other hand, explicit restrictions are calculated from (9) treating the α_i, β_{ij}, and γ_i as constants. The approximate budget share system (9) satisfies all the general restrictions with the exception of symmetry, iff $\gamma_i = \beta_{im}$. Then the system (9) reduces to (4):

$$w_i = \frac{\alpha_i + \sum_j \beta_{ij} z_j}{\alpha_m + \sum_j \beta_{mj} z_j}$$

It has symmetric substitution effects when equation (6) is satisfied. This will hold "globally" iff either $\beta_{ij} = \beta_{ji}$ (preferences given by (1)) or if preferences are given by (5).

Because (7) is homogeneous of degree one only at the base point, one cannot test simultaneously true and explicit restrictions from (9). However it could be argued that this is of little importance since the purpose of the theory is to test properties of true preferences at a point so that there is little reason to impose consistency of the approximation with utility maximization.

B. *Approximation of the Indirect Marginal Rates of Substitution*

The marginal rate of substitution (*MRS*) is another ordinal function which is of frequent use in demand analysis. Let f_{in} be the true *MRS* between commodities i and n, and approximate f_{in} to the first order around $(ln\,\bar{p}, ln\,\bar{y})$ by

$$(10) \quad \frac{w_i}{w_n} = \alpha_i + \sum_j \beta_{ij}\, ln\, p_j + \gamma_i\, ln\, y \qquad i = 1, \ldots, n-1$$

so that
$$\beta_{ij} = \frac{\partial(\bar{w}_i/\bar{w}_n)}{\partial\, ln\, p_j}, \quad \gamma_i = \frac{\partial(\bar{w}_i/\bar{w}_n)}{\partial\, ln\, p_j}$$

and
$$\alpha_i = \frac{\bar{w}_i}{\bar{w}_n} - \sum_j \beta_{ij}\, ln\, \bar{p}_j - \gamma_i\, ln\, \bar{y}$$
$$(i = 1, \ldots n-1; \quad j = 1, \ldots n)$$

Since $\Sigma_i w_i = 1$, we deduce from (10) the budget share system

$$(11) \quad w_i = \left[\alpha_i + \sum_j^n \beta_{ij}\, ln\, p_j + \gamma_i\, ln\, y\right] \div \left[1 + \sum_i^{n-1} \alpha_i + \sum_i^{n-1}\sum_j^n \beta_{ij}\, ln\, p_j + ln\, y \sum_i^{n-1} \gamma_i\right] \qquad (i = 1, \ldots n-1)$$

A test of both true and explicit homogeneity of degree zero is given by $\gamma_i = -\beta_{in}$ $(i \neq n)$. If homogeneity is imposed we obtain again the system (4):

$$w_i = \frac{\alpha_i + \sum_j^n \beta_{ij} z_j}{\alpha_m + \sum_j^n \beta_{mj} z_j}$$
$$i = 1, \ldots n;\ \beta_{ni} = 0,\ \gamma_n = 0,\ \alpha_n = 1$$

Then explicit restrictions are the same as in the expenditure approximation, whereas for true tastes to respect demand theory at the base point, one has to impose

$$\sum_i^{n-1} \alpha_i + \sum_i^{n-1}\sum_j^n \beta_{ij}\bar{z}_j = (1 - \bar{w}_n)/\bar{w}_n \qquad \text{(adding up)}$$

$$\beta_{ij} - \bar{w}_i\beta_{mj} - \bar{w}_j\beta_{im} = \beta_{ji} - \bar{w}_j\beta_{mi} - \bar{w}_i\beta_{jm} \qquad \text{(symmetry)}$$

The latter reduces to

$$\beta_{in} = \bar{w}_i\beta_{mn} + \bar{w}_n(\beta_{im} - \beta_{mi}) \qquad \text{for } i \text{ or } j = n$$

given adding up, homogeneity, and the defini-

tion of β_{ij}. A simple and equivalent expression for the symmetry condition in the *MRS* case:

$$\beta_{ij} - \frac{\overline{w}_j}{\overline{w}_n}\beta_{in} = \beta_{ji} - \frac{\overline{w}_i}{\overline{w}_n}\beta_{jn}$$

$$(i = 1, \ldots n-1; \quad j = 1, \ldots n)$$

can be derived from the symmetry of the "indirect Antonelli" functions

$$\frac{\partial(\overline{w}_i/\overline{w}_n)}{\partial z_j} - \frac{\overline{w}_i}{\overline{w}_n}\frac{\partial(\overline{w}_j/\overline{w}_n)}{\partial z_n}$$

We can conclude with the answer to question 4:

There are at least three different approaches to selecting a flexible functional form with which one can test properties of true preferences at a given point: approximation based on the utility function; the expenditure function; and the marginal rate of substitution functions. All of these lead to the same estimating model. However they differ in the parameter restrictions which correspond to the different properties of true preferences; hence any test of properties of true preferences must be conditional on the method of approximation. Since the approximation methods are all first order in the demand functions, choice between them cannot be based on their accuracy.

The utility approximation is implicitly cardinal and gives difficulties in discriminating between true and explicit restrictions. The expenditure approximation has the property that true preferences and preferences represented by the approximation cannot both satisfy the postulates of demand theory. This leaves the *MRS* approximation: it is ordinal; applicable to either the direct or indirect functions; and can satisfy properties of both true and approximating preferences.

IV. Econometric Methodology

Christensen, Jorgenson, and Lau proceed to estimate the approximate model without further restrictions than those emanating from the structure of preferences. In fact it is possible to use the properties of the approximation process in formulating the stochastic specification. In any of the approximations, the true budget share system can be written

$$(12) \qquad w_{it} = g_i(z_t) + \epsilon_{it}$$

and the approximate model

$$(13) \qquad w_{it} = g_i^A(z_t; \alpha, \beta) + \nu_{it}$$

where $\epsilon_t = [\epsilon_{it}]$ is a vector of random errors satisfying $E[\epsilon_t] = 0$, $\Sigma_i\epsilon_{it} = 0$, $E[\epsilon_t\epsilon_t'] = \Omega$ of rank $n-1$ and $E[\epsilon_t\epsilon_s] = 0$ $(t \neq s)$; while $\nu_t = [\nu_{it}]$ is a vector of random errors composed of ϵ_t and the approximation errors such that $\Sigma_i\nu_{it} = 0$. By construction

$$(14) \qquad \nu_{it} = \epsilon_{it} + \Delta g_i(z_t)$$

where $\Delta g_i(z_t) = g_i(z_t) - g_i^A(z_t; \alpha, \beta)$.

Now, an exact representation of (14) at any point z_t is

$$(15) \quad \nu_{it} = \epsilon_{it} + \Delta g_i(\overline{z}) + \sum_j \frac{\partial \Delta g_i(\overline{z})}{\partial z_j}(z_{jt} - \overline{z}_j) + \frac{1}{2}\sum_k\sum_j \frac{\partial^2 \Delta g_i(z^*)}{\partial z_k \partial z_j} \cdot (z_{kt} - \overline{z}_k)(z_{jt} - \overline{z}_j)$$

where $z_t^* = \theta_t\overline{z} + (1-\theta_t)z_t$ for $0 \leq \theta_t \leq 1$.

But for any of the approximation procedures used, the deterministic parts of the true and approximate models satisfy

$$(16a) \quad g_i(\overline{z}) = g_i^A(\overline{z}; \alpha, \beta) \qquad (\text{i.e., } \Delta g_i(\overline{z}) = 0)$$

$$(16b) \quad \frac{\partial g_i(\overline{z})}{\partial z_j} = \frac{\partial g_i^A(\overline{z}; \alpha, \beta)}{\partial z_j} \qquad \left(\text{i.e., } \frac{\partial \Delta g_i(\overline{z})}{\partial z_j} = 0\right)$$

Then, since $\overline{z}$ can be equated to zero by a suitable choice of units, (15) reduced to

$$(17) \quad \nu_{it} = \epsilon_{it} + \frac{1}{2}\sum_k\sum_j \frac{\partial^2 \Delta g_i(z^*)}{\partial z_k \partial z_j} z_{kt}z_{jt}$$

The a priori information on (17) is that

$$\sum_i \sum_k \sum_j \frac{\partial^2 \Delta g_i(z^*)}{\partial z_k \partial z_j} z_{kt} z_{jt} = 0$$

and moreover that

$$\sum_k \sum_j \frac{\partial^2 \Delta g_i(z^*)}{\partial z_k \partial z_j} = 0$$

iff g_i and g_i^A coincide everywhere.

The Hessian of Δg_i involves third-order errors of the relevant utility function or second-order errors of the demand functions. Neither economic theory nor the approximation procedure provide any further information about its behavior. Nevertheless it is clear that in general $E[\nu_t] \neq 0$ and $E[\nu_t z_t'] \neq 0$ unless true budget shares have the form of (4). Hence estimation of (13) by quasi-maximum likelihood with the usual normality assumption will lead to biased and inconsistent parameter estimates. The problem is then to find some means of incorporating the approximation error into the estimating form. Some simplifying hypothesis about the behavior of the approximation errors is clearly necessary to permit its estimation; otherwise (13) would not be identified since the approximation errors vary with time.

A possible approach is to proceed to estimate (13) with

$$(18) \qquad \nu_{it} = \epsilon_{it} + k_i \| z_t \|^2$$

where $\| z_t \|$ is the euclidean norm. This assumption has at least the local merit of validity for any k_i when z_t is sufficiently close to $\bar{z}$.

Explicit restrictions correspond to the case $g_i = g_i^A$ identically in z and hence $k_i = 0$. Tests of these restrictions will therefore always be conducted under the hypothesis $k_i = 0$.

On the other hand, if budget shares are not generated by the translog system (4), then $k_i \neq 0$. In this case, the relevant test is for the existence of true, but not translog, preferences at the point $\bar{z}$. Notice that including an estimate of the approximation error in the model permits the distinction of the two hypotheses that indirect translog preferences are globally valid and that an unknown indirect utility function exists at $\bar{z}$ which is approximated by the indirect translog. If $k_i = 0$ then $\beta_{ij} = \beta_{ji}$ tests the hypothesis that preferences are indeed translog. On the other hand if $k_i \neq 0$, then $\beta_{ij} = \beta_{ji}$ tests the existence of true unknown preferences at $\bar{z}$ approximated by the indirect translog.

In testing true restrictions, care must also be taken to formulate the relevant hypothesis correctly. In particular, both the constrained and the unconstrained models must satisfy (16). For $\bar{z} = 0$, this yields an unobservable constraint on α (except if $\bar{\epsilon}_i$ is not random and actually known). Similarly, one may restrict both models to have an identical approximation error or may allow the latter to vary between models. This problem requires further examination; in the empirical application which follows we choose the pragmatic course of permitting α_i and k_i to vary between models.

In conclusion, the answer to question 5 is:

The proposed methodology implies that estimation and testing of true tastes are conditional on the base point, the method of approximation used, and the assumptions made concerning the behavior of the approximation error.

V. Empirical Illustration

The preceding methodology has been applied to test the validity of demand theory Details of the data, the method of estimation, and the parameter estimates are given in the Appendix. A summary of the restrictions is given in Table 1.

1. First, the existence of an approximation error is tested in each of three systems: (i) with no further constraint than the system (9); (ii) with the constraint that budget shares be homogenous of degree zero in prices and income; (iii) with the twin constraints of homogeneity of degree zero and $\beta_{ij} = \beta_{ji}$. The purpose here is to test whether (9) is an acceptable functional specification for budget shares. Note that of these tests only the last is for the existence of explicit translog preferences. A direct test that preferences are given by (5a), (5b), or (5c) could have been but was not performed here. The results for these tests are given in Table 2.

2. Second, conditional on $k_i = 0$, tests

TABLE 1—SUMMARY OF RESTRICTIONS

	Indirect Utility ($\gamma_i = -\beta_{im}$)	Expenditures	Marginal Rate of Substitution
Homogeneity			
True	–	$\gamma_i = \bar{y}\bar{w}_i - \beta_{im}$	$\gamma_i = -\beta_{im}$
Explicit	–	$\gamma_i = -\beta_{im}$	$\gamma_i = -\beta_{im}$
Adding up			
True	–	$\alpha_m + \Sigma\beta_{mj} \ln \bar{p}_j + \gamma_m \ln \bar{y} = \bar{y}$	$\bar{w}_n = [\alpha_m + \Sigma\beta_{mj} \ln \bar{p}_j + \gamma_m \ln \bar{y}]^{-1}$ $\alpha_n = 1, \beta_{nj} = 0, \gamma_n = 0$
Explicit	–	–	$w_n = [\alpha_m + \Sigma\beta_{mj} \ln p_j + \gamma_m \ln y]^{-1}$
Engel and Cournot[a]			
True	–	$\gamma_m = \bar{y}\beta_{mj} = 0$	–
Explicit	–	–	–
Symmetry[a]			
True	β_{ij}	$\beta_{ij} - \bar{w}_j\beta_{im}$	$\beta_{ij} - \dfrac{\bar{w}_j}{\bar{w}_n}\beta_{in}$
Explicit	β_{ij}	$\beta_{ij} - w_j\beta_{im} - w_i\beta_{mj}$	$\beta_{ij} - w_j\beta_{im} - w_i\beta_{mj}$

Note $w_i = \dfrac{\alpha_i + \Sigma\beta_{ij} \ln p_j + \gamma_i \ln y}{\alpha_m + \Sigma\beta_{mj} \ln p_j + \gamma_m \ln y}$

[a]Given homogeneity and adding up.

are performed of the hypothesis that the explicit budget share system is homogenous of degree zero and that it has been generated by a translog indirect utility function ($\beta_{ij} = \beta_{ji}$). The purpose here is to test the validity of indirect translog preferences given the correctness of (9). The results for these tests are given in Table 3.

3. Third, conditional on $k_i \neq 0$, tests of the existence of true preferences at the base point are performed in each of the three approximations used: approximation of the indirect utility function; the expenditure equations; or the marginal rate of substitution functions. For each case, both true homogeneity of degree zero and symmetry of the substitution effects at the base point are tested. The results of these tests are shown in Table 4.

The empirical evidence in Tables 2–4 indicates that the functional specification (9) is rejected. Even so, conditional on (9) being the correct functional specification, explicit homogeneity of degree zero of the budget share system is rejected, but conditional on (4) being the correct specification, explicit symmetry of the Hessian of the indirect translog utility function is not rejected. The conclusion from Tables 2 and 3 would then be that the

TABLE 2—TESTS OF FUNCTIONAL SPECIFICATION[a]

Given Adding Up	(equation 9)	9.65[b]
Given Homogeneity	(equation 4)	13.52[b]
Given Translog	(equation 4 with $\beta_{ij} = \beta_{ji}$)	9.93[b]

[a]Twice the difference of the logarithm of the likelihoods.
[b]A value significant at the 1 percent level.

TABLE 3—TESTS OF EXPLICIT PREFERENCES[a]

Homogeneity	51.15[b]
Translog (given homogeneity)	3.94

[a,b]See Table 2.

TABLE 4—TESTS OF TRUE PREFERENCES[a]

	Approximation of Indirect Utility	Expenditure Function	MRS
Homogeneity (given adding up)	47.28[b]	20.70[b]	43.41[b]
Symmetry (given homogeneity and adding up)	7.54	33.13[b]	1.64

[a,b]See Table 2.

particular specification (9) and homogeneity of degree zero of this system, so that it reduces to (4), are unacceptable but that subsequent imposition of translog preferences involves only a minor further loss of explanatory power.

Table 4 indicates that with all three approximations true homogeneity of degree zero at the base point is decisively rejected. On the other hand, symmetry of the true substitution effects at the point is not rejected for either the indirect utility or marginal rate of substitution approximations although it is for the expenditure equation approximation. One would then conclude that results as to the validity of demand theory depend on the particular approximation made and that homogeneity of degree zero is a less acceptable hypothesis than the further hypothesis of symmetry of the substitution effects.

VI. Conclusion

At the theoretical level it must be realized that C-J-L's proposed methodology for testing the hypothesis of constrained utility maximization gives results that are conditional on the approximation method. Moreover it yields difficulties in distinguishing properties of true and approximate preferences. Nevertheless the translog system is important in its own right; once the problem of imposing the second-order conditions is resolved, it is likely to be of wide use since it can encompass a fairly broad range of consumer behavior.

At the empirical level, the results can be altered with variations in the method of approximation. However there is some evidence that homogeneity of degree zero of the budget shares causes more empirical problems than symmetry of the substitution effects. Of course our econometric approach outlined above has limitations:

1) The specification of the approximation error (18) may have been inappropriate. Other assumptions on ν_{it} or an alternative choice of norm could have been used.

2) Tests of true restrictions are conditional on the choice of the base point $\bar{z}$. At present the sensitivity of test results to variations in $\bar{z}$ is not known.

3) An interesting (but computationally expensive procedure) would be to test true restrictions on a subset of the data close enough to the base point and see the robustness of the results when this sample is progressively enlarged. In this case the arbitrariness of (18) becomes less important.

4) One should remember that a three-goods case imposes restrictions on the signs of the cross-substitution effects. To permit comparison of the different approximation methods with C-J-L's results, the same commodity classification and sample have been used in the empirical application (except for the omission of the war years). However it seems worthwhile to test the restrictions on a larger number of goods and to study the possible correlation between the number of commodities and the acceptance of a particular restriction.

APPENDIX

The data is taken from the *Survey of Current Business* and gives current and constant price expenditure data for durables, nondurables, and services over the period 1929–72. The price base is 1958 so p_i (1958) = 1 and expenditure series are normalized to give $y(1958) = 1$. The war years 1942–45 have been excluded because of their dissimilar nature.

For each model there is a vector of additive random errors $(\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{nt})$ satisfying $\Sigma_i \epsilon_{it} = 0$ (compare Section IV). The subvector $(\epsilon_{1t}, \epsilon_{it}, \ldots, \epsilon_{n-1t})$ is assumed to be normally distributed with zero mean, constant covariance matrix, and to be serially independent. Under these conditions estimates are obtained by the familiar procedure of deleting one equation and maximizing the likelihood of the remaining subsystem with respect to the parameters. In this empirical illustration, $n = 3$, and the third equation has been deleted.

Table 5 gives details of parameter estimates with t-ratios in parentheses. Column 1 represents unrestricted estimates (only adding up has been imposed). Column 2 gives

TABLE 5—PARAMETER ESTIMATES

	Indirect Utility			Expenditures			Marginal Rate of Substitution			Explicit Preferences		
	1	2	3	1	2	3	1	2	3	1	2	3
α_1	.142 (58.4)	.142 (68.3)	.141 (75.4)	.149 (54.9)	.148 (51.4)	.147 (145.2)	.373 (53.3)	.381 (65.2)	.381 (65.4)	.141 (60.8)	.140 (79.0)	.139 (83.8)
α_2	.477 (179.)	.484 (156.)	.488 (160.)	.482 (107.8)	.483 (83.5)	.487 (56.9)	1.252 (101.)	1.324 (79.2)	1.324 (77.2)	.475 (173.)	.486 (140.)	.484 (175.)
α_3	.382	.374	.371	.369	.368	.366	1.	1.	1.	.384	.374	.377
β_{11}	-.190 (1.30)	-.655 (3.00)	-.018 (.458)	-.087 (5.39)	-.059 (4.98)	-.024 (2.14)	.084 (.521)	-.043 (.263)	.052 (.369)	.079 (.854)	-.025 (.180)	.012 (.331)
β_{12}	.222 (1.27)	.493 (2.10)	-.024 (.527)	.006 (.451)	-.012 (1.09)	-.007 (.590)	.084 (.448)	.300 (1.79)	.195 (1.37)	-.084 (.592)	-.016 (.102)	-.029 (.801)
β_{13}	.010 (.113)	.220 (1.98)	-.069 (2.74)	.072 (6.11)	.073 (5.80)	.323 (2.81)	-.509 (4.94)	-.389 (4.95)	-.408 (5.38)	-.124 (1.71)	-.113 (2.08)	-.088 (7.74)
β_{21}	-.641 (1.05)	-2.509 (2.89)	-.024	-.256 (9.61)	.156 (6.52)	-.007	.554 (1.72)	-.587 (1.12)	.260	.369 (.908)	-.091 (.158)	-.029
β_{22}	1.020 (1.40)	2.349 (2.41)	.262 (2.89)	-.029 (1.35)	.037 (1.64)	.066 (2.03)	.564 (1.51)	2.637 (4.81)	1.969 (9.36)	-.223 (.336)	.476 (.648)	.331 (4.08)
β_{23}	-.164 (.508)	.945 (2.07)	-.281 (6.75)	-.193 (9.87)	-.197 (7.77)	-.051 (2.29)	-2.729 (14.1)	-1.454 (5.93)	-1.699 (10.7)	-.671 (2.17)	-.467 (2.53)	-.283 (6.26)
β_{31}	-.542 (1.33)	-1.490 (2.65)	-.069	-.169	-.096	.031	0.	0.	0.	-.022 (.071)	-.303 (.666)	-.088
β_{32}	.500 (1.06)	.838 (1.38)	-.281	.023	-.024	-.059	0.	0.	0.	-.184 (.393)	-.020 (.037)	-.283
β_{33}	.536 (2.76)	.921 (3.29)	.195 (2.91)	.121	.123	.019	0.	0.	0.	.278 (1.48)	.158 (1.13)	.296 (6.20)
γ_1	-.007 (1.05)	-.058	.111	.153 (35.7)	.147	.146	.268 (2.82)	.133	.160	.127 (1.48)	.123	.105
γ_2	-.222 (.499)	-.786	.044	.465 (65.8)	.487	.478	.827 (4.52)	-.597	-.530	.346 (.918)	.082	-.019
γ_3	-.307 (.927)	-.269	.155	.382	.366	.375	0.	0.	0.	.007 (.025)	.125	.008
k_1	-.014 (1.15)	-.017 (1.38)	-.009 (.739)	-.050 (8.14)	-.046 (7.31)	-.050 (7.04)	-.016 (1.53)	-.016 (1.58)	-.017 (1.70)	0.	0.	0.
k_2	-.020 (1.08)	.053 (3.05)	-.024 (1.36)	-.017 (1.66)	-.029 (2.31)	-.012 (.628)	-.002 (.126)	-.023 (1.33)	-.005 (.412)	0.	0.	0.
k_3	.035	.069	.032	.066	.076	.062	.018	.038	.021	0.	0.	0.

homogeneity restricted estimates at the base point (globally for the explicit case). Column 3 represents both homogeneity and symmetry constrained estimates at the base point (globally for the explicit case corresponding to the indirect translog).

REFERENCES

A. P. Barten, "The System of Consumer Demand Functions Approach: A Review," *Econometrica*, Jan. 1977, *45*, 23–52.

A. Brown and A. Deaton, "Surveys in Applied Economics: Models of Consumer Behavior," *Econ. J.*, Dec. 1972, *82*, 1145–236.

L. R. Christensen, D. W. Jorgenson, and L. J. Lau, "Transcendental Logarithmic Utility Functions," *Amer. Econ. Rev.*, June 1975, *65*, 367–83.

D. W. Jorgenson and L. J. Lau, "The Structure of Consumer Preferences," *Annals Econ. Soc. Measure.*, Winter 1975, *4*, 49–101.

L. J. Lau, "Applications of Duality Theory: Comments," in Michael D. Intriligator and David A. Kendrick, eds., *Frontiers of Quantitative Economics*, Amsterdam 1974.

***Survey of Current Business*,** "The National Income and Product Accounts of the United States, 1929–74," Statistical Tables.

Shadow Pricing in Distorted Economies

By TRENT J. BERTRAND*

Project evaluations often must be carried out in economies where prices are distorted from world prices by tariffs, taxes, or subsidies. If it is assumed that these price interferences cannot or will not be eliminated, it is then necessary to derive an appropriate "shadow" or accounting price system and to understand its properties.[1]

Recent work appears to lay a sound theoretical basis for shadow price analysis.[2] From the acceptance of world market prices as the shadow prices for traded products, this work provides a methodology for defining shadow factor prices. Its properties make possible well-defined judgements on social profitability, lead to a priori theoretical restrictions on results, and allow consistency checks on the nature of adjustments to market prices required to obtain shadow prices. Specifically, it has been shown that the shadow factor prices (a) are unique; (b) are stationary for wide variations in factor requirements in projects under analysis and/or for wide variation in factor endowments in the distorted economy (see Bhagwati and Wan); (c) imply borderline social profitability in all existing activities (see Diamond and Mirrlees, and Srinivasan and Bhagwati); and (d) are subject to constraints dependent on the nature of adjustments for other market factor prices—in particular if one shadow price exceeds its corresponding market price then some other shadow price must be below its corresponding market price (Diamond and Mirrlees). These theorems thus not only provide a theoretical basis for shadow pricing but lead to some astonishing theoretical constraints on the results.

The present paper shows that these theorems are sensitive to restrictions on the product factor-dimensions of the model and to assumptions concerning patterns of specialization. The analysis is then extended to models in which these restrictions are relaxed It is argued that the model with the number of products (n) in excess of the number of factors (m) provides the most useful framework for shadow pricing analyses. Within this model, all the results identified above collapse and alternative theorems must be established. The following results are obtained:[3] (a) shadow factor prices are indeterminate; (b) no theoretical limitations exist on feasible shadow factor prices—it is in general possible to define a feasible factor price for each and every factor anywhere from $+\infty$ to $-\infty$; (c) borderline social profitability occurs in at most m of the n single output production activities and may not occur in any; (d) there are no theoretical limitations on which existing industries will be socially profitable, unprofitable, or on the borderline of social profitability—it is in general possible to define sets of feasible shadow factor prices that will generate all three results for any particular activity; (e) all shadow factor prices may be above or below all market factor prices; and (f) shadow factor prices will not in general be stationary with changes in factor endowments in the distorted economy. While some of these results correspond to

*Professor of economics, department of economics, State University of New York-Binghamton. Helpful comments have been received from the managing editor and referee.

[1]Throughout this paper, we will treat price interferences due to taxes or subsidies as distortions between actual prices and the optimal price system If the analysis is expanded to take account of government objectives underlying price interventions, this may not be the case In my 1974 paper I have argued that accounting for government objectives may eliminate the need for adjustments to market prices.

[2]See, in particular, Ronald Findlay and Stanislaw Wellisz, T. N. Srinivasan and Jagdish Bhagwati, and Bhagwati and Henry Wan. Although not explictly set in the same theoretical framework, Peter Diamond and James Mirrlees generate results for constant cost industries which apply in the production system characterized by the "nonsubstitution theorem" emphasized by these other writers. The stimulus for much of this work was provided by I. M. D. Little and Mirrlees.

[3]Results (a) and (f) have been recognized in the literature. See Bhagwati and Wan.

extreme outcomes, they are useful in highlighting some inadequately appreciated difficulties with the theory of shadow pricing.

I. A Trade Theory Model of Shadow Pricing

I start with a standard model of the type used in recent contributions (see Findlay and Wellisz, Srinivasan and Bhagwati, and Bhagwati and Wan) where an attempt has been made to apply trade theory to the shadow pricing problem. Essential parts of the analysis are first introduced in a 2 x 2 model, with fixed world prices. This is shown in equations (1)–(7).

$$(1)\qquad \overline{X}_1 = L_1 f_1(k_1)$$

$$(2)\qquad \overline{X}_2 = L_2 f_2(k_2)$$

$$(3)\qquad p_1^*(1+t_1)(f_1 - f_1'k_1) = p_2^*(1+t_2)(f_2 - f_2'k_2) = w_L$$

$$(4)\qquad p_1^*(1+t_1)f_1' = p_2^*(1+t_2)f_2' = w_K$$

$$(5)\qquad L_1k_1 + L_2k_2 = K$$

$$(6)\qquad L_1 + L_2 = L$$

$$(7)\qquad p_1^*X_1 + p_2^*X_2 = p_1^*\overline{X}_1 + p_2^*\overline{X}_2 = \pi$$

Equations (1) and (2) specify the production functions f_i, satisfying the usual neoclassical conditions where k_i is the capital-labor employment ratio in the ith sector, L_i is the labor input in the ith sector, and $\overline{X}_i$ is the output of the ith sector. Equations (3) and (4) are equilibrium factor-market conditions showing equality of value marginal products between sectors with p_i^* denoting the fixed world price of commodity i, t_i denoting the fixed percentage distortion between domestic and world prices, f_i' denoting the partial derivative of f with respect to k_i, and $w_K(w_L)$ denoting the market price of a unit of capital (labor); (5) and (6) are full-employment conditions, where K and L denote the endowments of capital and labor; and (7) is the budget constraint for the economy showing that aggregate consumption X_1 and X_2, valued at world prices, equals aggregate output valued at world prices (denoted as π).

The constant world price and constant distortion assumptions imply that domestic prices are fixed in the analysis. Assuming no factor intensity reversals, it is well known from the factor price equalization literature that this uniquely determines domestic market factor prices and therefore efficient factor-employment ratios if both commodities are domestically produced, that is, if the factor-endowment ray falls in the cone of diversification (see John Chipman).

The above model can be used to derive a set of shadow prices (denoted as w^*) which define the opportunity cost of factors measured at world prices. Solving for this w^* price for, say, capital (denoted as w_K^*), we obtain

$$(8)\qquad w_K^* = \frac{d\pi}{dK} = p_1^*\frac{d\overline{X}_1}{dK} + p_2^*\frac{d\overline{X}_2}{dK} = \frac{p_1^* f_1 - p_2^* f_2}{k_1 - k_2}$$

where use has been made of equations (1)–(2) and (3)–(6) to define output and factor-employment changes with the withdrawal of one unit of capital. In recent papers (see Bhagwati and Wan, and Srinivasan and Bhagwati), the w^* factor prices have been proposed as shadow prices to be used in project evaluation analysis in economies with distortions.

The interpretation of (8) is straightforward in light of the Rybcyznski theorem on the effects of changes in factor supplies with fixed product factor prices and therefore fixed factor-employment ratios. Withdrawal of capital requires an increase of employment (and output) in the labor-intensive sector and a decrease of employment (and output) in the capital-intensive sector. The world price opportunity cost of withdrawing capital is then given by the difference in the average productivity of labor in the sector losing labor minus the average productivity of labor in the sector gaining labor at world prices ($p_1^* f_1 - p_2^* f_2$) multiplied by the shift in labor that occurs per unit of capital withdrawn (which can be shown by use of (5) and (6) to equal $1/(k_1 - k_2)$).

It should be noted that the w^* factor price may be negative. Since the average product of labor at domestic prices is always higher in the capital-intensive sector, the shift of labor must result in a fall in aggregate output

evaluated at domestic prices. However, if relative prices are sufficiently distorted, it is possible that the average product is as high or higher in the labor-intensive sector when evaluated at world prices. If this is the case, the numerator in the last line of (8) and the w^* factor price will be zero or negative.[4]

A second property of w^*, emphasized by Bhagwati and Wan, is that the w^* factor prices are stationary as long as factor changes in the distorted economy leave the factor-endowment ray in the Chipman cone of diversification. Within this nonspecialization region, market factor prices are constant in a fixed-product price model with as many products as there are factors. But with constant factor prices, all the terms in (8) defining the w^* factor price are also constant. This stationarity property would be an attractive feature of shadow factor prices justifying use of these prices even for large-scale projects as long as the diversification condition holds. However, this depends on both the acceptability of w^* price as an appropriate shadow price and the robustness of this "stationarity" property outside the narrow restrictions of the 2 x 2 model.

The w^* factor prices can be put in perspective with a more general set of shadow prices defined in terms of real income losses with withdrawal of factors of production from alternative uses by augmenting model (1)–(7) with equations (9)–(15):

$$W = W(U_A, U_B) \tag{9}$$

$$U^A = U^A(x_1^A, x_2^A) \tag{10}$$

$$U^B = U^B(x_1^B, x_2^B) \tag{11}$$

$$\frac{U_1^A}{p_1^*(1+t_1)} = \frac{U_2^A}{p_2^*(1+t_2)} = \lambda^A \tag{12}$$

$$\frac{U_1^B}{p_1^*(1+t_1)} = \frac{U_2^B}{p_2^*(1+t_2)} = \lambda^B \tag{13}$$

$$X_1 = x_1^A + x_1^B \tag{14}$$

$$X_2 = x_2^A + x_2^B \tag{15}$$

Equation (9) defines social welfare W as a function of the utility of two individuals, U^A and U^B; (10) and (11) show individual utilities to be functions of their own consumption of the two commodities; (12) and (13) are equilibrium conditions for utility-maximizing consumers facing tariff-distorted domestic prices; λ denotes the marginal utility of income, U_1 and U_2 are the marginal utilities from consumption of the X_1 and X_2 commodities, and the superscript denotes the individual; and (14) and (15) show aggregate consumption as the sum of individual's consumption.

Using equations (9)–(13) it is possible to define shadow factor prices in terms of the change in social welfare as a factor of production is withdrawn from the distorted economy. For instance, we define the shadow price of capital measured in units of social welfare $\tilde{w}_K^s$ as

$$\tilde{w}_K^s = \frac{\partial W}{\partial K} = \frac{\partial W}{\partial U^A}\lambda^A\Big(\sum_i p_i^*(1+t_i)\frac{dx_i^A}{dK}\Big) + \frac{\partial W}{\partial U^B}\lambda^B\Big(\sum_i p_i^*(1+t_i)\frac{dx_i^B}{dK}\Big) \tag{16}$$

If it is assumed that $(\partial W/\partial U^A)\lambda^A = (\partial W/\partial U^B)\lambda^B$, it is possible to divide through by this common marginal social utility of income to define a money measure of the change in social welfare. We denote such shadow prices as w^s with (16) accordingly reduced to

$$w_K^s = \sum_i p_i^*(1+t_i)\frac{dX_i}{dK} \tag{17}$$

$$= w_K + \sum_i p_i^* t_i \frac{dX_i}{dK} - \sum_i p_i^* t_i \frac{d\bar{X}_i}{dK} \tag{18}$$

$$= w_K^* + \sum_i p_i^* t_i \frac{dX_i}{dK} \tag{19}$$

where w_K in (18) is the market factor price for capital.

There are two possible assumptions involved in going from (16) to (17). First, it might be assumed that the social welfare

[4]This possibility is related to Harry Johnson's work on immiserizing growth in distorted economies. He showed that increasing the supply of a factor in a distorted economy could lead to a fall in real income. The conditions under which this would occur have been examined by Frank Flatters and myself. However, the implications of reversing the analysis to define the real costs of withdrawing a factor from the distorted economy for use in a new project were not recognized in our work.

function is known and that costless redistributions of income are carried out so as to maintain a socially optimal distribution of income—a social optimum characterized by $(\partial W/\partial U^A)\lambda^A = (\partial W/\partial U^B)\lambda^B$. Alternatively, it might be assumed that the social welfare function is blind to distributional changes and that $(\partial W/\partial U^A)\lambda^A = (\partial W/\partial U^B)\lambda^B$ regardless of the nature of the income distribution. Neither of these assumptions is particularly attractive, but one or the other is embedded in much of applied welfare economics.

Equation (17) can be manipulated to obtain (18) where the shadow price is shown to equal the market price adjusted for supply and demand effects in markets where market prices do not accurately define benefits or costs at the margin. Further manipulation of (18) allows the w^s shadow price to be related to the w^* shadow price discussed above as in (19). From (19), it is seen that w^s equals the opportunity cost evaluated at world prices w^*, plus the benefits or costs resulting from consumption changes in an economy with consumer price distortions. The w^* factor price taken by itself thus neglects the secondary costs or benefits from consumption changes in a distorted economy.

The neglect of consumption distortions in shadow pricing of factors should not, however, be interpreted in isolation from shadow pricing of outputs from projects being evaluated. As noted in the introduction, world prices are assumed to equal shadow prices for traded commodities in the methodology under consideration. However, a unit change in the availability of a traded commodity, say $\bar{X}_T$, reflected in a unit relaxation of the budget constraint (7), can be shown to result in a change in social welfare equal to p^s_T (measured in money terms on the same assumptions used in going from (16) to (18):

$$(20) \qquad p^s_T = p^*_T + \sum_i p^*_i t_i \frac{dX_i}{d\bar{X}_T}$$

That is, it equals the world price plus the benefits or costs resulting from consumption changes in distorted markets. Using the world prices as shadow prices of traded goods therefore also neglects consumption distortions.

Despite the neglect of the costs or benefits with consumption changes in distorted markets for both factor and product prices, the $w^* : p^*$ price system may be an appropriate methodology under certain conditions. For instance, suppose that the consumption effects of factor or output changes operate solely through changes in consumption possibilities at world prices (denoted by π as in (7)); that is, we have aggregate consumption functions of the form

$$(21) \quad X_i = X_i(\pi, p^*_j(1 + t_j)) \qquad i, j = 1, 2$$

with, in our model, all prices and distortions held constant. It would thus be possible to rewrite (19) and (20) as

$$(22) \qquad w^s_K = w^*_K(1 + \sum_i p^*_i t_i m_i) = w^*_K b$$

$$(23) \qquad p^s_1 = p^*_1(1 + \sum_i p^*_i t_i m_i) = p^*_1 b$$

where m_i is the marginal propensity to consume the ith commodity ($= \partial X_i/\partial \pi$). Since this proportionality between the $w^s : p^s$ and $w^* \cdot p^*$ price systems will hold for all factors and prices, the $w^* \cdot p^*$ price system will result in project evaluations with the same sign and ranking as those obtained with the $w^s \cdot p^s$ system as long as $b > 0$ in (22) and (23). A sufficient condition for this is that $m_i > 0$ for all i with $t_i > 0$. If this restriction were not imposed, a project that expands consumption possibilities at world prices could so bias consumption against the good where private benefits understate social benefits that society is worse off, that is, the two price systems could give conflicting project evaluations.[5] Thus, the $w^* : p^*$ price system either neglects consumption distortions or requires restrictive assumptions on the form and properties of the aggregate consumption function.

Since the $w^* : p^*$ price system relies on

[5]The model (7), (9) (15), and (21) can be solved for the marginal effect on social welfare of an extra unit of foreign exchange (which eases the budget constraint) to yield the proportionality factor b between the $w^s \cdot p^s$ and $w^* \cdot p^*$ price systems. The b term may therefore be interpreted as a shadow foreign exchange rate. However, unless aggregate consumption is a function of π, project-specific shadow exchange rates would have to be calculated to account for the specific use of foreign exchange earned or lost with particular projects.

world prices and the opportunity costs of factors evaluated in world prices, I shall term the procedure of using these product and factor prices the "world price methodology." The remainder of this paper is concerned with an evaluation of this methodology.

II. The World Price Methodology: The $n \times n$ Case

The w^* shadow factor prices in the 2 x 2 case generalizes to the $n \times n$ case as

$$[w^*] = [\frac{\partial \overline{X}}{\partial L}] [p^*] \tag{24}$$

where $[w^*]$ is a $n \times 1$ vector of shadow factor prices, $[\partial \overline{X}/\partial L]$ is a $n \times n$ matrix with the element in the hth row and ith column equal to $\partial \overline{X}_i/\partial L_h$, and $[p^*]$ is a $n \times 1$ vector of world prices.

By the full-employment conditions in the $n \times n$ model, we also have

$$\underset{n \times n}{[l]} \ \underset{n \times 1}{[X]} = \underset{n \times 1}{[L]} \tag{25}$$

where $[l]$ is a $n \times n$ matrix of input-output coefficients, $[X]$ is a $n \times 1$ vector of outputs, and $[L]$ is a $n \times 1$ vector of factor endowments.

With production of all n commodities in competitive equilibrium, we have

$$[l'] [w] = [p] \tag{26}$$

where $[w]$ is a vector of market wage rates, $[l']$ is the transpose of the $[l]$, and $[p]$ is a vector of market prices in foreign currency units (i.e., world prices multiplied by the trade taxes or subsidies on the particular commodity). World prices and the price interferences define $[p]$ and, by (26), $[w]$ if all n commodities are produced and if the inverse of $[l']$ exists. Assuming both these conditions hold, market factor prices and input-output coefficients are constant if world prices and price interferences are fixed and the economy remains within the same "cone of diversification." With coefficients in the $[l]$ matrix constant, we have from (25),

$$[\frac{\partial X}{\partial L}] = [l^{-1}]' \tag{27}$$

so that, by (24), we have an alternative definition of the w^* prices;

$$[w^*] = [l^{-1}]'[p^*] \tag{28}$$

The usefulness in defining the $w^* : p^*$ relationship in this way is twofold. First, it allows direct calculation of the shadow price vector from data on world prices and input-output relationships in the distorted economy *when the $n \times n$ dimensionality restrictions holds* in a way that is not obvious from the treatment in Section I. Secondly, it provides a somewhat startling result: *all existing activities in the distorted economy will be evaluated as on the borderline of social profitability by the world price methodology.* To state this in an alternative way following Srinivasan and Bhagwati, the ratio of the value of domestic factor inputs to the value of net output of traded goods using the $w^* : p^*$ price system (the domestic resource cost (*DRC*) ratios) will be unity for all existing activities. This equality of benefits and costs may (with use of (28)) be written as

$$[l'][w^*] = [p^*] \tag{29}$$

While it has received considerable attention, the economics of this proposition is somewhat trivial. In contrast to the $n > m$ case (where, as we shall see, the proposition collapses), factor endowments in the $n \times n$ model uniquely determine output of all commodities. Thus, resources withdrawn from the economy can only be used in exactly the same way as before if full employment is to be maintained and no new activities are to be introduced. The result amounts to the recognition that if the shadow costs of withdrawing resources from an existing activity are measured by the output losses valued at world prices, putting these resources back into these activities will yield zero profits by definition if the same world prices are used to value the resulting outputs.

The possibility of a negative w^* shadow factor price noted above in the discussion of equation (8) might suggest a counterexample. Changes in outputs along the Rybcyznski line which release the factor with the negative w^* factor price would increase aggregate output and consumption possibilities evaluated at world prices. How can it then be said that the

existing activities are both on the borderline of social profitability when reducing one and expanding the other in the way required to release the factor with the negative w^* leads to gains even if the withdrawn factor is left unemployed? However, a *new* activity has been introduced and the proposition holds only with *existing* activities. The new activity is the process of using the negative w^* factor to produce zero output—a zero-cost disposal activity. While this removes the apparent inconsistency, it also indicates the limited usefulness of the proposition.

Note also that if the $[p]$ and $[p^*]$ price vectors differ, the $[w]$ and $[w^*]$ price vectors will also differ and the cost-minimizing production technique with the w^* prices would differ. Thus, a withdrawal of resources from an existing industry for use in a public enterprise in producing the same products, with the shadow cost-minimizing techniques along the same production *function*, would be socially profitable. However, this again represents a new activity (even though it is an existing industry using existing technology) and therefore again violates the severe restrictions on which the "borderline of social profitability of existing activities theorem" rests.

The theorems on shadow factor prices and the relationship between the w^* and w price system can be depicted in the Lerner factor-space diagram. For purposes of simplifying the diagram, the units of measurement of commodities are chosen so that all world product prices are unity. The unit isoquants for commodity $\bar{X}_2$ and $\bar{X}_1$, plotted in Figure 1, represent the same value of output *at world prices*. Domestic prices, however, may diverge from these world prices due to trade taxes or subsidies. It is assumed, to simplify the diagram, that no trade interferences exist for the $\bar{X}_2$ commodity but that production of commodity $\bar{X}_1$ is protected either through an import tariff, production subsidy, or export subsidy. The isoquant for $p_2^*/p_1^*(1 + t_1)$ units of $\bar{X}_1$ and the unit isoquant of $\bar{X}_2$ are of equal value at *domestic prices*. With both commodities produced in the domestic economy, the corresponding costs of production must be identical at market factor prices with producers choosing the cost-minimizing techniques. This requires the iso-cost line defining

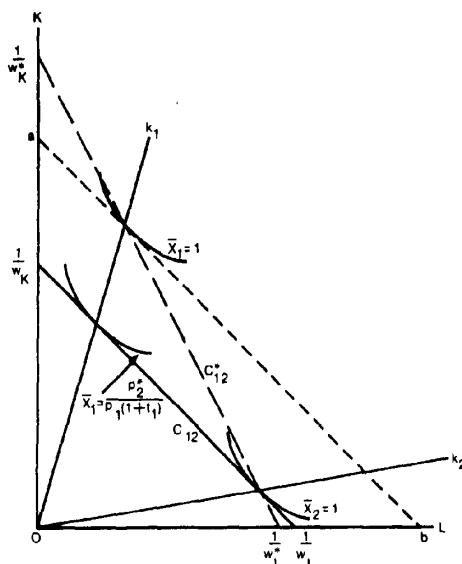



FIGURE 1

production costs to be tangent to the equal valued isoquants. As is known from work on the factor-price equalization theorem, this iso-cost line is unique if factor-intensity reversals do not occur; it is shown in Figure 1 as C_{12}. Since C_{12} is tangent to the unit isoquant for $\bar{X}_2$ and p_2^* is equal to one by assumption,

$$(30) \qquad C_{12} = w_K K + w_L L = p_2^* \bar{X}_2 = 1$$

The intersections of the C_{12} line with the axes of the factor-space diagram define reciprocals of real factor prices for capital and labor in terms of commodity X_2 (chosen as the numeraire because its price is not subject to distortions).[6] With these unique factor prices, cost-minimizing capital-labor ratios are defined by k_1 and k_2. If the capital labor endowment of the economy is in the k_1 0 k_2 cone (Chipman's cone of diversification), then the two outputs will be produced in the unique ratios required for full employment. Although not drawn to avoid complicating the diagram, this could be shown by completing a parallelogram between the origin and the factor-endowment

[6]This can be seen by letting L in (30) equal zero and seeing that the corresponding K, the intersection of C_{12} with the k axis, equals $1/w_K$.

point using the k_1 and k_2 factor-employment rays. The ratio of the sides of the parallelogram would measure the ratio of outputs required for full employment.

With the input-output coefficients defined by the k_1 and k_2 factor-employment ratios, (30) shows that unit production costs evaluated at the w^* prices must equal the p^* prices (i.e., "borderline social profitability for both activities" must hold). The iso-cost line, defining shadow factor costs so that this condition is satisfied, is uniquely determined as C^*_{12} in Figure 1. With C^*_{12}, shadow production costs of a unit of X_1 equal production costs of a unit of X_2 (equal valued at world prices), relative w^* factor prices are defined by the slope of C^*_{12}, and the reciprocal of real w^* factor prices are defined by the intersection of C^*_{12} with the axes of the diagram. The Lerner factor-space diagram can therefore also be used to define the world price methodology shadow prices.

The normalization on the unit isoquant for $\bar{X}_2$ with the resulting rotation of the C^*_{12} iso-cost line relative to C_{12} iso-cost line establishes that an *upward* adjustment in the market price for labor ($w^*_L > w_L$ in Figure 1) *must* result in a corresponding *downward* adjustment in the other factor price $w_K > w^*_K$). In the confines of the 2 x 2 model, this qualitative interconnection between adjustment to market factor prices provides a useful consistency check on applied social profitability analysis.

Note that the only result dependent on the economy's factor endowment point (as long as it remains in the cone of diversification) concerns the ratio of outputs required for full employment. This means in particular that both the w and w^* prices defined in Figure 1 are constant for all variations in the factor endowment within the $k_1 0 k_2$ cone. This stationarity property has been emphasized by Bhagwati and Wan. It is an attractive aspect of the solution since the calculated shadow factor prices will hold regardless of the size of the project under consideration as long as a diversified production structure is maintained.

All of the propositions discussed in the context of the n x n model appear to provide insights into properties of shadow prices that suggest guidelines for estimating, adjusting, or interpreting shadow prices in a distorted economy. However, it will be argued that each of these properties (uniqueness, zero social profitability of existing activities, predictable interconnection between adjustments to market factor prices required to obtain shadow prices, and stationarity) are all based on dimensionality restrictions on the model so severe that the results are of little relevance to practical problems of shadow pricing.

Most of the difficulties arise in the $n > m$ case. However, several theoretical problems are evident in the 2 x 2 example when specialization occurs: (a) factor prices are characterized by extreme sensitivity to the nonspecialization assumption reflected in the possibility of discrete jumps in the w^* prices, (b) these jumps may represent an increase or decrease that cannot be predicted solely on the basis of the factor which is being withdrawn from the distorted economy; and (c) both shadow factor prices in the 2 x 2 model could be greater or less than the corresponding market prices in real terms.

The first property is evident in Figure 1 when a withdrawal of capital results in specialization in the labor-intensive $\bar{X}_2$ sector Since in this example there are no trade restrictions on commodity $\bar{X}_2$, there are no effective distortions in production activities[7] if the economy is specialized in $\bar{X}_2$ and w^* shadow factor prices equal the market factor prices. A discrete shift therefore occurs in shadow factor prices from w^*_K and w^*_L defined by C^*_{12} with diversification to the w_K and w_L factor prices defined by C_{12} when the factor-endowment ratio equals k_2. In this particular example, the shadow price of capital shifts upwards at the specialization point and then increases in line with the marginal productivity of capital in the $\bar{X}_2$ sector with further withdrawal of capital. However, an example based on a taxed rather than protected X_1 sector would result in a discrete fall in the shadow price of capital at the specialization

[7]The treatment of consumption distortions in the world price methodology is discussed in Section I.

point. The shadow price ruling in the diversification region therefore cannot be judged a priori as an underestimate or overestimate of the shadow price with specialization. The same point can be made with respect to the shadow price of labor if so much labor were removed that the economy shifted to a factor endowment ratio equal to k_1 implying specialization in $\overline{X}_1$. The shadow price of the factors in the world price methodology then equals the marginal products in $\overline{X}_1$ production valued at world prices. The slope of line ab defines the relative marginal products at the unit isoquant for $\overline{X}_1$ and is parallel to the C_{12} iso-cost line. The intersection of this line with the axes defines point a, the reciprocal of the marginal product of capital, and point b, the reciprocal of the marginal product of labor, both in units of either commodity 1 or 2 since by assumption $p_1^* = p_2^* = 1$. The shadow factor prices with specialization in $\overline{X}_1$ are therefore defined as $1/a$ for capital and $1/b$ for labor. Compared to shadow prices ruling in the diversification region, there is a discrete fall in the shadow price of labor (even though the shift from the diversification region to the specialization point results from a rise in the residual capital-labor endowment) and a discrete increase in the shadow price of capital.[8]

The possibility of the required adjustment to market factor prices being *in the same direction* is also illustrated at the point of specialization in $\overline{X}_1$. Market factor prices are defined by the C_{12} line as w_L and w_K. Shadow factor prices are defined by $1/b$ and $1/a$, respectively. Relative market and shadow prices are identical, a necessity when factors are drawn from a single sector. However, the market prices reflect marginal products at the inflated domestic prices while the shadow prices value marginal products at world prices. The relationship between shadow and market prices is given as

$$(31)\quad w_h^* = w_h - p_1^* t_1 \frac{\partial \overline{X}_1}{\partial L_n} = w_h/(1+t_1) \qquad h = 1,2$$

by use of (18) and (19). Neglecting consumption distortions, this does not mean that an economy specialized in the production of $\overline{X}_2$ with price distortions is worse off than an economy with the same production pattern without price distortions since the tariff revenue is redistributed. However, it does mean that the market prices overvalue both labor and capital drawn from the protected sector for potential use in alternative projects.

Discrete changes in factor prices in directions not predictable solely on the basis of factor use or on the nature of market price adjustments for other factors complicates shadow price analysis outside the diversification region. However, even within the diversification region, the properties of shadow prices identified above are on fragile ground.

III. The World Price Methodology: The $n > m$ Case

The "fixed price" programming result that the same number of commodities will be produced as there are factors justifies some special attention to the $n = m$ case. However, such a model is not necessarily the best framework of analysis. Assumptions of common technologies across countries and free trade may lead to factor-price equalization, implying that every commodity can be produced in every country even without trade restrictions; that is, world prices are not randomly given independent of production costs elsewhere. With greater relevance to the present problem, price interferences are often instituted so as to permit domestic viability for industries not competitive at world standards and to tax sectors efficient by world

[8]An alternative illustration of these discrete jumps in shadow factor prices is available in the product space diagram. The shadow price in the world price methodology defines the *shift* in the consumption possibilities line as production adjusts with the withdrawal of the corresponding factor. The production point shifts along a negatively sloped Rybcyznski line until specialization and then along the axes towards the origin after specialization. The discontinuous jumps in the shadow price reflects the kink in the locus of production points at specialization. This diagramatic apparatus is not developed here since it does not as easily generalize into the many product world where our attention will be concentrated.

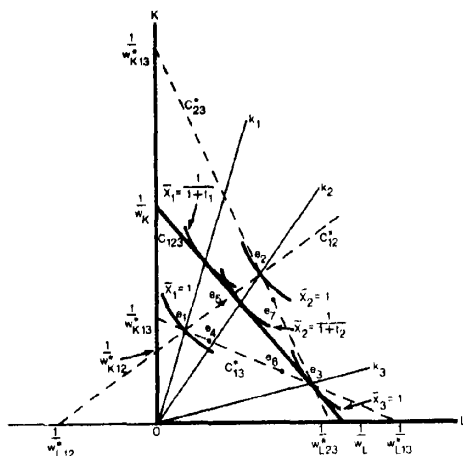



FIGURE 2

standards. This means that domestic market prices are biased towards consistency with diversified economic structures. On both these grounds, there is considerable appeal to the $n > m$ framework. This also represents the model dimension used in virtually all emprical work in this area.

Since all points to be made in this section hold for the two-factor three-product model, the Lerner factor-space diagram can be used with one commodity added. The unit isoquants for $\overline{X}_1$, $\overline{X}_2$, and $\overline{X}_3$ are plotted with units of measurement again chosen so that world prices are all equal to unity. In the case shown in Figure 2, $\overline{X}_2$ would never be produced without distortions to world prices. Regardless of factor endowments, it would always be more profitable to produce some combination of $\overline{X}_1$ and $\overline{X}_3$ and trade for $\overline{X}_2$. Assume, however, that in the distorted economy the producer price for $\overline{X}_2$ is raised by a positive distortion t_2 (either a production subsidy, import tariff, or export subsidy) and the producer price for $\overline{X}_1$ is reduced (by either an import subsidy, production tax, or export tax) so that domestically $1/(1 + t_1)$ units of $\overline{X}_1$ and $1/(1 + t_2)$ units of $\overline{X}_2$ are of equal value as one unit of $\overline{X}_3$ (which is not subject to any price distortion and which serves as the numéraire). If all three commodities are produced, the market production costs of these three outputs must be identical with cost-minimizing production techniques chosen by producers. This condition is satisfied with the iso-cost line C_{123} which, as in Figure 1, defines the relative market factor prices by its slope and the reciprocals of the absolute factor prices in terms of the $\overline{X}_3$ commodity by its intersections with the axes. The cost-minimizing factor proportions with domestic market prices are k_1, k_2, and k_3 with $k_1 0 k_3$ defining Chipman's cone of diversification. If the factor endowment is in this cone, factor prices will be defined by C_{123} and variation of the capital-labor endowment will affect the output mix but not the market factor price system. The actual outputs must be such that the full-employment condition is satisfied with the k_1, k_2, and k_3 factor-employment ratios in the three industries. This does not, contrary to the $n \times n$ case, uniquely determine the structure of outputs, and this has important implications for the w^* shadow factor prices.[9] This point and its implications will be developed in the context of the general $n > m$ case. The 3 x 2 model will then be used to illustrate the results.

The full-employment conditions in the $n \times m$ model yield

$$(32) \qquad \underset{m \times n}{[l]} \; \underset{n \times 1}{[X]} = \underset{m \times 1}{[L]}$$

with matrices as defined in (25) generalized beyond the same number of products as factors restriction. Note that the parameters in $[l]$ are constants within the cone of diversification since market factor prices remain constant. However, with $n > m$, $[l^{-1}]$ is not defined and a unique solution for outputs does not exist.[10] Thus the output changes in the economy resulting from a withdrawal of a factor from the distorted economy cannot be uniquely defined and (32) therefore cannot be used to uniquely define the w^* shadow factor prices.

This is a crucial result that is easily

[9]The nature of this indeterminacy and the implications for the production possibilities in the economy is explored in the paper by Jaroslav Vanek and myself.

[10]This indeterminacy can be easily visualized in the factor-space diagram. The factor-employment rays must be combined to pass through the factor-endowment point With more than two activities, the combination of rays satisfying this condition is not unique.

explained in the context of the adjustments to market wage rates required to obtain the w^* shadow prices. This adjustment depends on output changes due to a withdrawal of a factor from existing activities weighted by the price distortions. The price distortions are known, but it is not possible to uniquely determine the output changes, and therefore the adjustments to market prices are indeterminate.

The indeterminacy of w^* shadow prices can be removed by adding $n - m$ restrictions on output changes to the model. For instance, if outputs were held fixed in sectors $m + 1$ through n, we would have

$$(33) \qquad \underset{m \times m}{\left[\frac{\partial X}{\partial L}\right]} = \underset{m \times m}{{}^{v}[l^{-1}]'}$$

where ${}^{v}[l]_{m \times m}$ is $[l]_{m \times n}$ with the last $m - n$ columns deleted. Note (33) is derived from the full-employment conditions when the last $(m - n)$ outputs are held fixed. By the definition of the w^* shadow factor prices, we have

$$(34) \qquad \underset{m \times 1}{w^*} = \underset{m \times m}{{}^{v}[l^{-1}]'}\ \underset{m \times 1}{{}^{v}[p^*]}$$

where ${}^{v}[p^*]$ denotes $[p^*]$ with the prices for the outputs that are held constant deleted. By premultiplying (33) by ${}^{v}[l]'$, it is seen that borderline social profitability (or a unit domestic resource cost ratio) is guaranteed for the m activities not subject to the fixed output constraints, that is, the activities involved in the resource reallocation process if factors are drawn from the distorted economy. If ${}^{v}[l]_{m \times m}$ is of full rank m, the $\Sigma_h l_{hi} w_h^* = p_i^*$ borderline social profitability condition can hold for at most m activities[11] and therefore social profitability must be either positive or negative in all activities where outputs have been held fixed. While this provides a generalization of sorts to the borderline of social profitability for all existing activities theorem in the $n \times n$ model, its collapse as any guideline in shadow pricing is complete once it is realized that there are $\binom{n}{m}$ combinations of the fixed output restrictions, $\binom{n}{m}$ sets of w^* factor prices built on this restriction, and therefore $\binom{n}{m}$ combinations of m activities with zero social profitability.

Based on specifying restrictions to eliminate the production indeterminacy in the $n > m$ model, it might seem that it is possible to define the range of feasible solutions for the w^* shadow factor prices. This, it might be argued, would allow upper and lower limits to be placed on the shadow costs of any new activity so that, if the range of shadow factor prices was not too great, the indeterminacy problem would not prevent conclusions on social profitability from being drawn. This argument however is not correct as in general no restriction whatsoever can be placed at the theoretical level on any w^* shadow factor price in the diversification region. In general, each shadow factor price can range from $+\infty$ to $-\infty$. Corresponding to each of the $\binom{n}{m}$ fixed output restrictions there is a unique w^* for each factor and the range of these w^*'s can certainly be defined. However, the production indeterminacy can be eliminated by defining m activities used in deriving the w^* prices from linear combinations of the individual production activities. This can be done in an infinite number of ways and any shadow price can be generated for any factor. This is efficiently illustrated with the Lerner factor-space diagram.

Recall that by choice of units for the three goods, the unit iso-quants in Figure 2 are equal valued at world prices. When the output indeterminacy is eliminated by fixing output of $\overline{X}_i$, the shadow factor prices are defined by the zero profitability condition for the other two activities as in (33). This condition is satisfied by C_{12}^* when $\overline{X}_3$ is held constant, C_{13}^* when $\overline{X}_2$ is held constant, and C_{23}^* when $\overline{X}_1$ is held constant. The intersections of these shadow iso-cost lines with the axes again define the reciprocals of shadow factor prices in real terms using world prices. The zero social profitability condition, while satisfied for the activities not subject to the constant output restriction, does not hold for the other

[11]This full rank condition rules out the factor input point (determined by constant market prices) in any activity being a linear combination of the similarly defined factor-input points producing equal value output in other activities. This will be interpreted below for the 3 x 2 case in terms of the Lerner factor-space diagram.

activity. For instance, the shadow factor prices defined by C_{13}^* show that the $\overline{X}_2$ sector is socially unprofitable. Producing one unit of $\overline{X}_2$ at e_2 would necessarily require inputs of capital and labor that exceed the value of these inputs along C_{13}^* but would only yield equal valued output.

It is also possible to avoid the production indeterminacy by defining joint product activities as linear combinations of the single product activities. In an example aimed at illustrating the lack of feasible theoretical limits on shadow factor prices, we define a linear combination of activities for producing one unit of $\overline{X}_1$ and $\overline{X}_3$ (with factor-input points e_1 and e_3 in Figure 2) which results in a joint product activity producing both $\overline{X}_1$ and $\overline{X}_3$ using the factor inputs shown at e_4 in Figure 2; an output of equal value to one unit of $\overline{X}_1$ or one unit of $\overline{X}_2$. Consider the consequences of providing production determinacy by fixing the relative weights between $\overline{X}_1$ and $\overline{X}_3$ outputs as given at e_4. The resulting shadow factor prices are defined by a C^* iso-cost line through points e_2 and e_4 (not drawn to avoid cluttering the diagram). The intersections of this iso-cost line with the axes would define a very high negative shadow price of labor and a very high positive shadow price of capital. In fact these shadow prices can be made to approach plus ∞ by increasing the weight of X_3 in the joint activity shifting e_4 towards the $0k_2$ factor-employment line. The economics of the result is easy to understand; for example, a reduction of a unit of capital will lead to a large reduction in the joint e_4 activity and a large expansion in the e_2 activity in order to maintain full employment of the remaining factors. This involves transferring labor from the joint activity where the average product of labor is high to X_2 production where the average product is low, both evaluated at world prices. The factor withdrawal and reallocation results in a large decline in the value of output at world prices and is reflected in the high shadow price of capital.

Two further points might be noted, both self-evident in light of results already established. With suitable choice of the two activities used to define shadow factor prices (including joint activities), any shadow price can be established as theoretically possible spanning those approaching $+\infty$ or $-\infty$, those connected with the derivation in which one output is held fixed defined by the C^* shadow iso-cost lines drawn in Figure 2, and everything in between. Secondly, if any of the single product activities are not used separately in the m activity basis for the shadow factor price solution, social profitability in this activity will not be zero. Thus, borderline social profitability will occur in at most m single product existing activities and may not occur in any (the latter result occurring if only joint production activities are used in obtaining production determinacy).

In the $n \times n$ model, a theorem was established to the effect that if the shadow factor price for some factor was above its market price, the shadow price for some other factor price must be below its shadow price. This theorem does not survive in the $n > m$ model The simplest way to see this is to eliminate the production indeterminacy by choosing two linear combinations of the single product activities to form a basis for a shadow price solution in such a way as to define an iso-cost line which does not intersect the market price iso-cost line C_{123}. Such linear combinations are defined in Figure 2 as e_5 and e_6. These equal value activities are chosen so that the iso-cost line passing through these points (not drawn to avoid cluttering the diagram) is parallel to the C_{123} iso-cost line and results in shadow factor prices higher than the corresponding market factor prices.

The stationarity property in the $n > m$ model also collapses. The shadow factor prices connected with any particular set of restrictions used to eliminate the production indeterminacy are constant in the diversification region, but there is no reason to expect these restrictions to remain constant with changes in factor supplies and shadow factor prices can shift in an infinite number of ways.

Furthermore, any product activity can be shown to be socially unprofitable, on the borderline of social profitability, or socially unprofitable with a feasible set of shadow factor prices. This can be seen by noting that it is always possible to find two activities to be used in generating shadow factor prices (including joint activities) that will define

iso-cost lines that leave a particular single product activity on the resulting shadow iso-cost line or on the socially profitable or unprofitable side of it. The borderline case is obvious since including any single product activity in the basis for the factor price solution will result in zero social profitability as can be seen from (33). The rest of the proposition may be established by considering any activity in Figure 2 and confirming that shadow price sets exist which result in both social profitability and unprofitability. A good candidate is activity $\bar{X}_2$ which, as noted in introducing Figure 2, would never be a socially efficient activity in the nondistorted economy. If the two activities used in deriving the shadow factor prices are defined at e_6 (producing $\bar{X}_1$ and $\bar{X}_3$) and e_7 (producing $\bar{X}_2$ and $\bar{X}_3$), the corresponding shadow iso-cost line through these points (not drawn to avoid cluttering the diagram) will result in positive shadow factor prices for both commodities and will pass to the northeast of e_2. This means (since output at e_1, e_2, e_3, and any linear combination of these points is equal valued at world prices) that the value of a unit of $\bar{X}_2$ output equals the shadow cost of factors on the shadow iso-cost line through e_6 and e_7. Since production of $\bar{X}_2$ uses less labor and capital than some points on this iso-cost line, the $\bar{X}_2$ activity is socially profitable. This same procedure can be used to confirm that even if a particular activity uses less (more) of both factors than any other activity producing equal valued output, social unprofitability (profitability) is still feasible in a distorted economy.[12]

With these results, it is useful to sum up by noting that all four favorable properties on shadow factor prices established in the $n \times n$ model with diversification (uniqueness; zero social profitability of existing activities, predictable interconnection between adjustments to market factor prices required to obtain shadow prices; and stationarity) collapse in the $n > m$ framework. The results that do emerge (indeterminacy; no theoretical limitations on feasible factor shadow prices; borderline social profitability in at most m sectors and possibly in none; no theoretical limitations on which industries will be socially profitable, unprofitable, or on the boderline of profitability; possibilities of all shadow factor prices being above or below all market prices; lack of stationarity) have one thing in common. They tend to destroy rather than support any theoretical guidelines restricting shadow factor prices or their movements.

III. The World Price Methodology: The $n < m$ Case

The $n < m$ case is subject to the practical limitation that any attempt to apply results run into the immediate difficulty of deriving a sufficiently narrow classification of inputs to be consistent with the model. The $n < m$ case nevertheless does avoid the production indeterminacy, and shadow factor prices will be uniquely defined. This can be seen by noting that with fixed input-output coefficients, there are not enough distinct factor-employment techniques to guarantee that the full-employment conditions are satisfied. Changes in factor prices are therefore required, changes which lead to variation in the input-output coefficients until a full-employment solution is reached. However, since the fixed market factor prices no longer hold, complicated price effects involving substitution elasticities in production processes will be determinants of the shadow factor prices. None of the other conclusions noted in the introduction as providing theoretical guidelines or constraints on shadow prices survive in the $n < m$ case.

[12]The proof of these seemingly paradoxical propositions requires the use of joint activities which are partly based on the activity being evaluated. Thus, in proving $\bar{X}_2$ potentially profitable, a joint activity involving $\bar{X}_2$ and $\bar{X}_3$ has to be specified. Nevertheless, since one can think of drawing resources from a distorted economy for use in a particular activity in a way that affects existing production of that output, the thrust of the argument is not affected.

REFERENCES

T. Bertrand, "The Shadow Exchange Rate in an Economy with Trade Restrictions," *Oxford Econ Pap.*, July 1974, *26*, 185–91.

——— and F. Flatters, "Tariffs Capital Accumulation and Immiserizing Growth," *J.*

Int. Econ., Nov. 1971, *1*, 453–60.

J. Bhagwati and H. Wan, Jr., "The 'Stationarity' of Shadow Prices of Factors in Project Evaluation, with and without Distortions," *Amer. Econ. Rev.*, June 1979, *69*, 261–73.

J. S. Chipman, "A Survey of the Theory of International Trade: I," *Econometrica*, July 1965, *33*, 477–519; "II," Oct. 1965, *33*, 685-760; "III," Jan. 1966, *34*, 18–76.

P. A. Diamond and J. A. Mirrlees, "Private Constant Returns and Public Shadow Prices," *Rev. Econ. Stud.*, Feb. 1976, *43*, 41–48.

R. Findlay and S. Wellisz, "Project Evaluation, Shadow Prices, and Trade Policy," *J. Polit. Econ.*, June 1976, *84*, 543–52.

H. G. Johnson, "The Possibility of Income Losses from Increased Efficiency or Factor Accumulation in the Presence of Tariffs," *Econ. J.*, Mar. 1967, *77*, 151–54.

I. M. D. Little and James A. Mirrlees, *Manual of Industrial Project Analysis in Developing Countries*, Vol. 2, Paris 1969.

T. N. Srinivasan and J. Bhagwati, "Shadow Prices for Project Selection in the Presence of Distortions: Effective Rates of Protection and Domestic Resource Costs," *J Polit. Econ.*, Feb. 1978, *86*, 97–116.

J. Vanek and T. Bertrand, "Trade and Factor Prices in a Multi-Commodity World," in Jagdish Bhagwati et al., eds., *Trade, Balance of Payments, and Growth*, Amsterdam 1971.

A Theory of Federal Debt Management

By V. VANCE ROLEY*

The impact of federal debt management on the economy is an important issue in macro-economics that remains unresolved. Hypotheses concerning the impact of a shift from long- to short-term Treasury securities have included a full range of possibilities encompassing stimulative, restrictive, and neutral effects.[1] This range of comparative statics solutions derived from macro-economic models of financial market equilibrium evolves from different assumptions concerning the degree of substitutability among financial assets. In some cases, particular financial assets have been assumed to be perfect substitutes because of similar attributes, while in other cases it has been assumed that different pairs of financial assets are imperfect gross substitutes or even gross complements. Assumptions concerning the imperfect gross substitutability or complementarity of financial assets are represented by sign restrictions on the partial derivatives of aggregate financial asset demand equations. However, these sign restrictions may be questioned since the underlying asset demand analysis has typically suffered from at least two shortcomings. First, asset demands have not been explicitly derived from the optimizing behavior of individual economic agents. Second, an explicit role for uncertainty about future yields on financial assets has not been adequately developed.

The analytical framework used in this paper to investigate the impact of federal debt management may be contrasted to previous macro-economic theories of financial market equilibrium in several respects. First, individual economic agents are assumed to maximize their expected utility in the familiar manner developed by Harry Markowitz and Tobin (1958). Second, the equilibrium prices of different securities are determined in the capital asset pricing model framework introduced by William Sharpe and by John Lintner (1965) that typically has been applied only to equity valuation. Third, the aggregate demand equations for financial assets that emerge from the analysis suggest that assumptions concerning the gross substitutability or complementarity of financial assets are not important when assessing the qualitative impact of debt management. Instead, investors' assessments of the covariances of future random security returns determine this impact. Assumptions about covariances do not translate into unambiguous substitutability or complementarity relationships in aggregate demands.

An equilibrium model of the financial

*Senior staff economist, Council of Economic Advisors, on leave from the Federal Reserve Bank of Kansas City I am grateful to Peggy Brockschmidt for research assistance, and to G. O. Bierwag, Benjamin Friedman, Carl Gambs, Stephen Miller, and an anonymous referee of this *Review* for their comments and suggestions. The views expressed here are solely my own and do not necessarily represent the views of the Council of Economic Advisors or the Federal Reserve Bank of Kansas City.

[1]A stimulative impact on the economy is implied when investors aggregate securities according to time-to-maturity—for example, long-term Treasury securities with equity. This aggregation scheme is adopted by John Maynard Keynes (1930, 1936) and Axel Leijonhufvud. The possibility of a restrictive impact is, for example, implied by Earl Rolph. In Rolph's analysis, there is an optimal maturity composition of the federal debt which further implies that the substitution of short- for long-term Treasury securities could in some instances result in a restrictive impact. Debt management has no impact when short- and long-term Treasury securities are assumed to be perfect substitutes. For an analysis of this case, see James Tobin (1961, 1963). The inability of debt management to affect the Treasury security yield curve—which is consistent with the perfect substitutes hypothesis—has been supported by empirical research using reduced-form interest rate equations. A partial list of this empirical research includes Arthur Okun, Robert Scott, Franco Modigliani and Richard Sutch (1966, 1967), G. O. Bierwag and M. A. Grove; Michael Hamburger and William Silber. However, in my doctoral dissertation, I estimate a disaggregated structural model of the U.S. Treasury securities market and find that debt management is capable of affecting the Treasury security yield curve.

sector of the economy is presented in Section I of this paper.[2] *Using this model, expressions for the aggregate values of private and Treasury securities are derived in Sections II and III, respectively.* The implications of alternative assumptions concerning investors' assessments of the joint distribution of future security prices are compared with the comparative statics results obtained in previous macroeconomic models of financial sector equilibrium. In addition, selected sample statistics are contrasted to these alternative sets of assumptions. In Section IV, alternative federal debt management policies are evaluated in terms of their effects on the level of nonfinancial economic activity as measured by changes in the aggregate value of private securities. The particular link between the financial and nonfinancial sectors of the economy envisioned in this paper is based on the effects on aggregate investment from different levels of Tobin's (1969) "q."[3] Thus, the amount of private investment is affected through changes in the aggregate value of private securities, where the price of new capital goods is assumed to be constant. The main conclusions of this paper are summarized in the final section

I. The Model

A model of financial market equilibrium is employed to derive the expressions for the market valuation of private securities and Treasury debt. It is assumed that m investors choose among a riskless asset (money), a private security (equity), and two maturities of Treasury securities, with all securities treated as pure discount securities to simplify the analysis. It is also assumed that investors base their portfolio selection decisions on estimates of future nominal security prices which are independent of current market prices, and that investors' portfolio selection horizons do not exactly correspond to either of the two maturities of the Treasury securities. Hence, investors assume some risk when investing in Treasury securities.

In choosing the optimal portfolio, each investor i maximizes his expected utility U_i (which is assumed to be a function of the mean $E_i(W_{i1})$ and variance $V_i(W_{i1})$ of his random end-of-period wealth W_{i1}) subject to his wealth constraint. Thus, the optimal portfolio of the ith investor follows from the maximization of the Lagrangian function

$$(1)\quad L_i = U_i[E_i(W_{i1}), V_i(W_{i1})] + \lambda_i(\mathbf{P}_0'\mathbf{q}_i + m_i - W_{i0})$$

with respect to $\mathbf{q}_i$ and λ_i, where

- $\mathbf{q}_i$ = vector consisting of the number of units of equity (q_{iE}), short-term Treasury securities (q_{iS}), and long-term Treasury securities (q_{iL}) held by the ith investor
- $\mathbf{P}_0$ = vector consisting of the current market prices per unit of equity (P_E), short-term Treasury securities (P_S), and long-term Treasury securities (P_L)
- m_i = dollar amount invested in the riskless asset by the ith investor
- W_{i0} = initial wealth of the ith investor
- λ_i = Lagrange multiplier appropriate for the ith investor

It is additionally assumed that each investor's assessment of future random security prices may be described by the same joint probability distribution. Thus, the vector of means and the variance-covariance matrix of future security prices may be expressed as

$$E_i(\mathbf{P}_1) = \mu$$
$$V_i(\mathbf{P}_1) = \Sigma, \text{ with typical element } \sigma_{jk}\ (j, k = E, S, L)$$

where $\mathbf{P}_1$ = vector consisting of the random end-of-period prices per unit of equity (P_{E1}), short-term Treasury securities (P_{S1}), and long-term Treasury securities (P_{L1})

Given that the random end-of-period wealth

[2]However, this is a partial equilibrium model in the sense that nonfinancial economic variables remain at initial values throughout the analysis. For examples of the use of this methodology, see Tobin (1963, 1969).

[3]The effects on nonfinancial economic activity resulting from changes in liquidity are abstracted from in this paper. For a discussion concerning the liquidity impact of debt management, see Warren Smith.

of the ith investor may be represented as

$$W_{i1} = \mathbf{P}_1'\mathbf{q}_i + m_i$$

the corresponding mean and variance of random end-of-period wealth are

$$E_i(W_{i1}) = \mu'\mathbf{q}_i + m_i$$

$$V_i(W_{i1}) = \mathbf{q}_i'\Sigma\,\mathbf{q}_i$$

Therefore, from the first-order conditions of the constrained maximization problem (1), it follows that the ith investor's vector of demands is

$$(2) \qquad \mathbf{q}_i = \rho_i\Sigma^{-1}(\mu - \mathbf{P}_0)$$

where[4] $\rho_i = -(\partial U_i/\partial E_i)/2(\partial U_i/\partial V_i) > 0$

By summing over the m individual demands in (2), the market demands for the risky assets may be expressed as

$$(3) \qquad \mathbf{q}^D = \rho\Sigma^{-1}(\mu - \mathbf{P}_0)$$

where $\qquad \rho = \sum_{i=1}^{m} \rho_i$

Assuming that ρ is constant,[5] the equilibrium prices of securities may be derived using a market-clearing identity equating the vector of demands (2) with exogenous supplies

$$(4) \qquad \mathbf{q}^D = \mathbf{q}^0 = (q_E^0, q_S^0, q_L^0)'$$

This implies, from (3) and (4), that the vector of equilibrium prices is

$$(5) \qquad \mathbf{P}_0 = \mu - (1/\rho)\Sigma\,\mathbf{q}^0$$

Thus, the equilibrium price of security j is

$$P_j = \mu_j - (1/\rho)(\sigma_{jE}q_E^0 + \sigma_{jS}q_S^0 + \sigma_{jL}q_L^0)$$

which is equal to investors' common expectation of its future price minus an adjustment consisting of the product of the systematic risk component $(\sigma_{jE}q_E^0 + \sigma_{jS}q_S^0 + \sigma_{jL}q_L^0)$ and the market price of risk $(1/\rho)$.[6]

The market demands (3) and equilibrium prices (5) are analogous to those used either explicitly or implicitly in macro-economic models of financial market equilibrium. It is of particular interest to note that sign restrictions on the covariances $\sigma_{jk}(j, k = E, S, L)$ are sufficient to determine the directions of the impacts of changes in security supplies $q_j^0(j = E, S, L)$ on security prices $P_j(j = E, S, L)$. However, Oliver Blanchard and Mary Plantes have shown for asset demands similar to (3) that sign restrictions on these covariances are *not* sufficient to determine gross substitutability or complementarity properties among securities—that is, the signs of the elements of the Jacobian matrix $\partial\mathbf{q}^D/\partial\mathbf{P}_0$. Therefore, the investigation of the impact of federal debt management focuses on alternative assumptions about these covariances and not on the usual notions of gross substitutability or complementarity.

II. The Aggregate Market Value of Private Securities

The impact of federal debt management on the overall level of economic activity examined in Section IV involves the effect on aggregate investment from a change in the aggregate value of private securities for a given price of new capital goods. The crucial component of this link involves the role of relative supplies of Treasury securities in the determination of the aggregate market value of private securities. This relationship is derived and analyzed below.

Using (5), the aggregate value of private securities may be expressed as

[4]Following Tobin (1958), each investor i is assumed to be risk averse thereby implying $\rho_i > 0$. It is also implicitly assumed that each investor's demand for the riskless asset (money) satisfies a nonnegativity constraint.

[5]To explicitly derive equilibrium security prices, a variety of assumptions concerning the functional form of the ρ_i may be used. For example, in the quadratic utility case

$$U_i = E_i(W_{i1}) - b_i[E_i(W_{i1})]^2 - b_iV_i(W_{i1})$$

the individual ρ_i may be expressed as

$$\rho_i = [E_i(W_{i1}) - 2b_i]/2b_i$$

which results in equilibrium prices similar to those in (5), (see Jan Mossin). The case where the ρ_i are constant may be derived from the assumption of constant absolute risk aversion (see John Pratt and Kenneth Arrow) and joint normally distributed security prices. This case has been extensively examined by Lintner (1969, 1970) and is very useful in investigating the relationships among security supplies and prices.

[6]For a discussion concerning the effects of market size and investors' risk aversions on the market price of risk, see Lintner (1970).

$$(6)\quad V_E = \mu_E q_E^0 - (1/\rho)\sigma_{EE}(q_E^0)^2 - (1/\rho)\sigma_{ES} q_E^0 q_S^0 - (1/\rho)\sigma_{EL} q_E^0 q_L^0$$

The aggregate market valuation equation (6) indicates that the total value of private securities equals the usual expression which is valid in the absence of Treasury securities—that is, the risk-adjusted expected value in terms of the parameters of the distribution of the price of private securities—adjusted by terms involving the covariances of the prices of private and Treasury securities. These additional risk adjustment terms may be positive, negative, or zero, depending on investors' assessments of the covariances between future equity and Treasury security prices. The contribution of the variance (σ_{EE}) and covariances (σ_{ES}, σ_{EL}) in determining the market value of equity reflects the risk associated with uncertain future security prices. If the variance (σ_{EE}) increases, for example, the market value of equity decreases due to the increased riskiness involved in holding equity. Similar increases in the covariances (σ_{ES}, σ_{EL}) also result in a lower market value of equity because larger covariances imply that investors view the behavior of Treasury security and equity prices to be more alike. Thus, there is greater risk associated with holding a portfolio of risky assets which results in a lower market value of equity.

It may also be seen from (6) that the supplies of short- and long-term Treasury securities may be determinants of the aggregative value of private securities. The degree of their importance depends on the magnitudes of the covariances between the future price of equity and future prices of Treasury securities. To further analyze the role of Treasury securities in the valuation of private securities, and to facilitate the representation of (6) in $q_S - q_L$ space, the market valuation expression (6) may be rearranged as

$$(7)\quad q_S^0 = \rho(\mu_E/\sigma_{ES}) - (\sigma_{EE}/\sigma_{ES})q_E^0 - \rho(1/\sigma_{ES}q_E^0)V_E - (\sigma_{EL}/\sigma_{ES})q_L^0$$

Thus, holding the aggregate value and supply of private securities constant in (7), the marginal rate of substitution between short- and long-term Treasury securities is

$$(8)\quad \frac{dq_S^0}{dq_L^0} = -\frac{\sigma_{EL}}{\sigma_{ES}}$$

Hence, the manner in which Treasury securities may be substituted by the government authorities to keep the aggregate value of private securities constant depends on the signs and magnitudes of investors' assessments of the covariances of the future price of equity with the future prices of different maturities of Treasury securities.[7] Previous studies have, however, used a variety of explicit and implicit assumptions that translate into specific assumptions concerning these covariances. Thus, several cases involving alternative sets of assumptions concerning the signs of the covariances (σ_{ES}, σ_{EL}) are examined immediately below.

A. *The Implications of Alternative Covariance Assumptions*

To examine the role of the covariances (σ_{ES}, σ_{EL}) in determining the impact of a change in the size and/or composition of the stock of outstanding Treasury securities on the aggregate value of private securities, three cases involving alternative assumptions concerning the covariances are considered. First, for the case of positive covariances ($\sigma_{ES}, \sigma_{EL} > 0$), iso-value lines corresponding to three different levels of the aggregate value of private securities are illustrated in Figure 1, where $V_E^* > V_E^1 > V_E^0$.[8] It is apparent from (8) and Figure 1 that in order to leave the aggregate value of private securities constant when changing the composition of Treasury securities, short-term Treasury securities must be sold (purchased) and long-term Treasury securities purchased (sold) by the government authorities according to σ_{ES}/σ_{EL}. Furthermore, from (6), a decrease in q_S^0 or q_L^0 increases the aggregate value of private secu-

[7] Again, it should be noted that because all nonfinancial variables—including the price of new capital goods—remain constant throughout the analysis, the equilibrium properties of the model are those of a short-run equilibrium.

[8] The iso-value lines in Figure 1 correspond to different levels of the aggregate value of private securities (V_E) in equation (7).

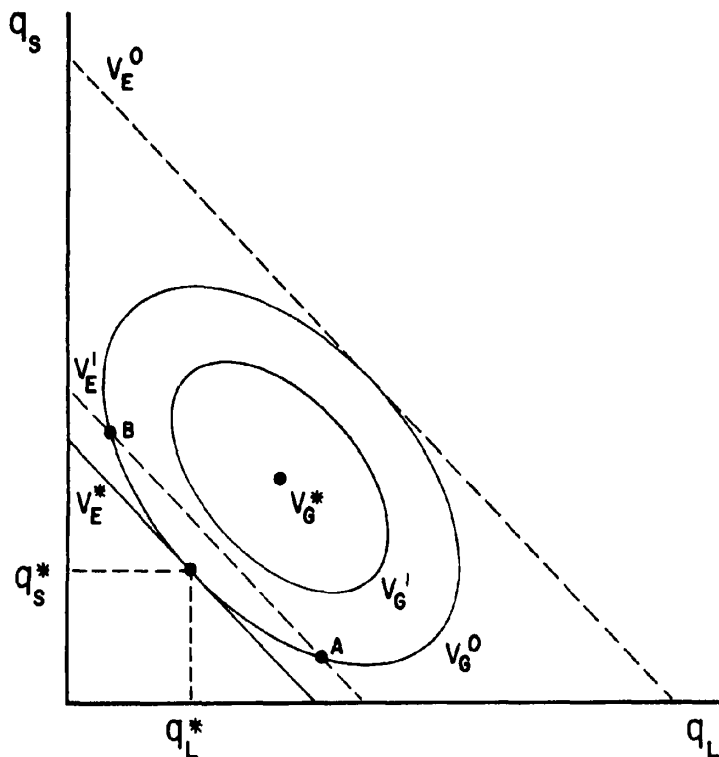



FIGURE 1

rities, *ceteris paribus*. This pair of covariance assumptions, therefore, has implications similar to those of the models presented by Rolph, Richard Musgrave, and Oswald Brownlee and Ira Scott, although these authors employed different methodologies.[9]

Second, for the case of negative covariances ($\sigma_{ES},\sigma_{El} < 0$), the iso-value lines in Figure 1 again may be used to represent different levels of the aggregate value of private securities, where $V_E^* < V_E^1 < V_E^0$. Furthermore, the rate of substitution between short- and long-term Treasury securities required to keep the aggregate value of private securities constant has the same sign as in the previous case. However, in this case, a decrease in short- or long-term Treasury securities decreases the aggregate value of private securities, *ceteris paribus*. The implications arising from this pair of covariance assumptions correspond to those presented by Tobin (1963). In Tobin's model, if the supply of long-term Treasury securities increases, then investors, at current market prices, hold too few short-term Treasury securities relative to long-term Treasury securities, and too few private securities relative to their aggregative holdings of Treasury securities. Thus, in equilibrium, the aggregate market value of private securities increases.[10] This result

[9]Brownlee and Scott, however, use a model similar to that developed herein. Nevertheless, their analysis differs in several respects. First, they assume that investors possess increasing absolute risk aversion. Second, they do not explicitly derive an expression for the aggregate value of private securities corresponding to (6). Finally, they do not explicitly derive the marginal rate of substitution between short- and long-term securities required to keep the aggregate value of private securities constant. Instead, they assert, without reference to a condition corresponding to equation (8), that $dq_S^0/dq_l^0 < 0$.

[10]It should be noted that Tobin (1963) bases these results on the risk relationship between Treasury securities and equity. However, he concludes that the correlations between future equity returns and future returns on

TABLE 1—SAMPLE STATISTICS FOR SELECTED SECURITY YIELD AND PRICE DATA

	Mean	Variance-Covariance Matrix		
Yields		*RTB*	*RTL*	*REP*
Three-Month Treasury Bills (*RTB*)	5.71	1.88	0.41	1.62
Long-Term Treasury Securities (*RTL*)	6.51	0.41	0.32	1.13
Standard and Poor's Earnings-Price Ratio (*REP*)	8.11	1.62	1.13	5.87
Prices		*PTB*	*PTL*	*PSI*
Three-Month Treasury Bills (*PTB*)	94.61	1.49	3.12	6.35
Long-Term Treasury Securities (*PTL*)	61.32	3.12	21.44	28.15
Standard and Poor's Common Stock Index (*PSI*)	95.92	6.35	28.15	126.83

Note: All yield and price data used in the computations, except the price of Treasury bills, are quarterly averages of monthly data published by the Board of Governors of the Federal Reserve System. The price of Treasury bills is computed from the market yield. Means, variances, and covariances are sample moments computed for the period beginning in 1970:I and ending in 1977:IV

corresponds to the movement from V_E^* to V_E^1 in Figure 1.

The final case considered involves a negative covariance between the prices of short-term Treasury securities and equity, and a positive covariance between long-term Treasury securities and equity ($\sigma_{ES} < 0, \sigma_{EL} > 0$). In this case, the iso-value lines representing different levels of the aggregate value of private securities have positive slopes. Thus, in order to leave the aggregate value of private securities constant and to change the composition of Treasury securities, both short- and long-term Treasury securities must be sold (purchased) by the government authorities according to σ_{ES}/σ_{EL}. Furthermore, from (6), either a decrease (increase) in short-term Treasury securities or an increase (decrease) in long-term Treasury securities reduces (increases) the aggregate value of private securities. Similar results also follow from the implicit frameworks used by Keynes (1930, 1936) and Leijonhufvud. In these studies, the primary distinction between securities is their length to maturity. Thus, if there is an increase in the supply of long-term Treasury securities, the prices of all long-term securities fall.

The three cases of alternative covariance assumptions considered above are those most frequently encountered in the literature. It may be noted that another case (i.e., $\sigma_{ES} = \sigma_{EL} = 0$) implies that a change in the composition of Treasury securities has no effect on the aggregate value of private securities. Thus, for a given supply of private securities, any point in $q_S - q_L$ space is consistent with the aggregate value of private securities that results.

It is also of interest to compare the three pairs of covariance assumptions with the actual historical behavior of Treasury security and equity prices.[11] Thus, in Table 1, sample means, variances, and covariances of yields and prices of selected types of securities are separately included. According to the numerical values reported, the positive covariance case ($\sigma_{ES}, \sigma_{EL} > 0$) is that which is consistent with the recent historical data.[12]

short- and long-term Treasury securities are small without specifying their signs. In order to derive the implications of his model in the framework presented herein, it must additionally be assumed that $\sigma_{ES}, \sigma_{EL} < 0$. In other models presented by Tobin (1961, 1969), the unambiguous result reported in the text does not necessarily follow.

[11]A further step in the analysis involves the assumption that historical observations of market yields or prices are used by investors to form assessments of the distribution of future security prices. In a paper by Benjamin Friedman and myself which compares three hypotheses (unitary, rational, or autoregressive) involving investors' expectations of future corporate bond holding-period yields, the hypothesis corresponding to this assumption of autoregressive expectations appears to be the most representative.

[12]It should also be noted that two additional sample periods for these data were investigated in the manner

This suggests that an increase in either short- or long-term Treasury securities depresses the aggregate value of private securities, and that short- and long-term Treasury securities may be substituted for one another at roughly a 4 to 1 ratio while keeping the aggregate value of private securities constant.

III. The Aggregate Market Value of Treasury Securities

The determinants of the aggregate market value of Treasury securities are examined in this section. It is shown that there exists a variety of different combinations of Treasury securities which have the same aggregate market value. This result is used in Section IV to describe the possible tradeoffs between short- and long-term Treasury securities available to the government authorities when implementing discretionary federal debt management policy.

Using (5), the aggregate value of Treasury securities may be expressed as

$$(9)\quad V_G = [\mu_S - (1/\rho)(\sigma_{ES}q_E^0)]q_S^0 + [\mu_L - (1/\rho)(\sigma_{EL}q_E^0)]q_L^0 - [(2/\rho)\sigma_{SL}]q_S^0q_L^0 - [(1/\rho)\sigma_{SS}](q_S^0)^2 - [(1/\rho)\sigma_{LL}](q_L^0)^2$$

which implies that a given aggregate market value of Treasury securities may be represented by an iso-value ellipse in $q_S - q_L$ space.[13] It is also apparent from (9) that the aggregate market value of Treasury securities depends on the supplies of Treasury and private securities, parameters of the joint distribution of future security prices, and the market price of risk $(1/\rho)$. Furthermore, the sign of the covariance between the future prices of short- and long-term Treasury security prices is of special importance when representing (9) in $q_S - q_L$ space. In particular, the sign of the covariance σ_{SL} determines the slope of the directrices of a given iso-value ellipse, and, therefore, the position of the ellipse itself. Alternative assumptions about the sign of the covariance σ_{SL} are considered below.

A. *The Implications of Alternative Covariance Assumptions*

Different values of the covariance (σ_{SL})—assuming the variances $(\sigma_{SS}, \sigma_{LL})$ are constant—represent different assessments by investors concerning the correlation of the future prices of short- and long-term Treasury securities. For the case of a positive covariance $(\sigma_{SL} > 0)$, three different levels of the aggregate value of Treasury securities are illustrated in Figure 1, where $V_G^* > V_G^1 > V_G^0$. The ellipses are drawn under the assumption that $\sigma_{LL} > \sigma_{SS}$, which, along with $\sigma_{SL} > 0$, corresponds to the numerical values in Table 1. The directrices of the ellipses are positive as a result of the positive covariance assumption.[14] Furthermore, it is generally true that an increase in the magnitude of the covariance increases the positive slope of a given directrix.[15] Hence, the higher the assessed correlation between future short- and long-term Treasury security prices, the "flatter" the ellipses illustrated in Figure 1.

Two additional properties of the market valuation equation (9) consistent with the positive covariance case as well as the others

---

ported in Table 1. The additional periods were 1960:I–1964:IV and 1965:I–1969:IV. In both cases, the yield data exhibited uniformly positive covariances. However, for the price data, the covariances between short-term Treasury securities and equity and between long-term Treasury securities and equity were negative.

[13]The discriminant of equation (9) is equal to

$$(1/\rho)^2\sigma_{SS}\sigma_{LL} - (1/\rho)^2\sigma_{SL}^2$$

which is positive except in the presence of perfect correlation. Thus, this equation represents either the empty set, a point, or an ellipse.

[14]If the coefficient on the mixed term $(q_S^0q_L^0)$ in the market valuation equation (9) (i.e., $(2/\rho)\sigma_{SL}$) is positive, then it can be shown that a given iso-value ellipse has positively sloped directrices.

[15]If b is the *negative* of the slope of a directrix, and e is the corresponding eccentricity (assumed to be constant), then

$$\frac{db}{d\sigma_{SL}} = \frac{-(1/\rho)(1+b^2)}{[(2/\rho)\sigma_{SL}b + e^2]}$$

which is negative if $[(2/\rho)\sigma_{SL}b + e^2] > 0$. Given reasonable and suitably small values of $(1/\rho)$ (see Lintner, 1970), this condition will generally be satisfied.

are of particular interest. First, for a given combination of Treasury securities on the portion of the ellipse that is convex to the origin, there exists a semi-ellipse with greater amounts of each security but with the same aggregate value. Second, there exists a combination of Treasury securities such that any change in either of the amounts of short- or long-term Treasury securities results in a decrease in aggregate market value. This point is represented by V_C^* in Figure 1. These properties are either additional or contradictory to those listed in previous research dealing with federal debt management (see, for example, Rolph; Brownlee and Scott).

The primary differences between the positive covariance case and the two other possible covariance assumptions ($\sigma_{SL} < 0, \sigma_{SL} = 0$) are reflected by the slopes of the directrices of the respective sets of iso-value ellipses.[16] In particular, the directrices of a set of iso-value ellipses are negative and zero for the negative and zero covariance cases, respectively. Thus, in contrast to the positive covariance case, there are generally smaller ranges in which the one-for-one substitution of short- and long-term Treasury securities by the government authorities result in only small changes in their aggregate value. The negative and zero covariance cases are not considered in detail in the investigation of alternative federal debt management policies that follows. Instead, primary emphasis is placed on the positive covariance case because of the empirical evidence presented in Table 1 and the observation that this is the assumption most commonly encountered in the literature (see Rolph; Brownlee and Scott; Tobin, 1963; Neil Wallace; William Nordhaus and Henry Wallich).

IV. Alternative Federal Debt Management Policies

Federal debt management policy is defined herein as the policy of government authorities concerning a change in the composition of a given size of the outstanding interest-bearing marketable Treasury debt held by private investors. Several features of this definition should especially be noted. First, since debt management operations involve purchasing and selling Treasury securities at market-determined prices, the size of the federal debt is defined in terms of total market value. This allows a more accurate description of the debt management possibilities facing government authorities than those in terms of total value at par. Second, the total size of the Treasury debt is defined to exclude monetary debt. Policy concerning the substitution of money for interest-bearing federal debt, or vice versa, is assumed to be in the province of monetary policy. Finally, the portion of the total outstanding Treasury debt relevant to debt management policy is the amount held by private investors. This follows from the familiar practice of consolidating the federal government's balance sheet. Thus, federal debt management is jointly determined by the actions of the Treasury and the Federal Reserve System.

Alternative federal debt management policies are defined in terms of different targeted levels of the aggregate value of private securities for a given aggregate value of Treasury debt. A particular debt management policy may correspond to the maximization of the aggregate value of private securities, or the attainment of a lesser value for purposes of reducing aggregate demand in the nonfinancial sector of the economy. Since federal debt management operations are in terms of the market values of Treasury securities, the possibilities facing the government authorities are assumed to be described by the iso-value ellipse corresponding to the current aggregate value of the federal debt.[17]

[16]Two illustrative properties may be conveniently derived for the zero covariance case First, the center of an iso-value ellipse under this assumption is

$$\frac{\mu_S - (1/\rho)(\sigma_{ES} q_E^0)}{(2/\rho)\sigma_{SS}}, \frac{\mu_L - (1/\rho)(\sigma_{EL} q_E^0)}{(2/\rho)\sigma_{LL}}$$

which is in the positive quadrant for most relevant cases. Second, the eccentricity of a given iso-value ellipse is

$$e = \frac{\sigma_{LL} - \sigma_{SS}}{\sigma_{LL}}$$

[17]Defining federal debt management possibilities in terms of a given iso-value ellipse does not necessarily imply an absence of costs incurred by the government authorities when changing the composition of the federal

Sufficient conditions that enable federal debt management to be effective in economic stabilization may be easily enumerated. First, at least one of the covariances between future prices of Treasury securities and equity is nonzero. This implies that a given aggregate value of private securities has a distinct representation in $q_S - q_L$ space, as in Figure 1. Second, short- and long-term Treasury securities are viewed as distinct types of securities by investors.[18] In the absence of this property, a one-for-one substitution of short- and long-term Treasury securities would leave their aggregate value unchanged as well as the aggregate value of private securities since the marginal rate of substitution along an iso-value line would also equal -1. Finally, private security prices affect the nonfinancial economy through the effect of Tobin's q on aggregate private investment. In the model, the value of existing physical capital is equated with the aggregate value of private securities, and the price of new capital goods is assumed to be constant. Thus, the aggregate value of private securities determines the level of q, with the higher the level of q, the greater the investment in new capital goods. This relationship implies that different iso-value lines in Figure 1 correspond to different levels of aggregate private investment.

A. *Stimulative Federal Debt Management Policies*

The federal debt management policy for maximum economic stimulation, with respect to the link investigated in the model, is the policy which maximizes the aggregate value of private securities for a given aggregate value of Treasury debt (V_G^0). Based on the signs of the covariances presented in Table 1 ($\sigma_{ES}, \sigma_{EL}, \sigma_{SL} > 0$), the optimal combination of short- and long-term Treasury securities (q_S^*, q_L^*) occurs at the tangency of the iso-value ellipse V_G^0 and the iso-value line V_E^* in Figure 1. If the optimum is an interior solution, then[19]

$$\frac{P_L - (\mu_L - P_L)}{P_S - (\mu_S - P_S)} = \frac{\sigma_{EL}}{\sigma_{ES}} \tag{10}$$

That is, the optimum occurs at the point where the ratio of short- and long-term Treasury security prices minus their corresponding expected returns is equal to the ratio of the covariances of the future prices of short- and long-term Treasury securities with the future price of equity, respectively.[20] Furthermore, if the composition of the federal debt corresponds to point A in Figure 1—which implies that the left-hand side of (10) is less than the right-hand side—the government authorities should sell short-term securities and purchase long-term securities to stimulate the economy. The opposite holds for stimulative debt management policies when the outstanding federal debt corresponds to point B in Figure 1.

The condition for the optimal combination of Treasury securities (10) is also consistent with the tangency V_E^0 with V_G^0 in Figure 1.[21]

---

debt. For small changes in the composition of the federal debt, if the slope of an iso-value ellipse at the existing composition of the federal debt equals the ratio of the prices of long- and short-term Treasury securities, with the government authorities being able to carry out their debt management operation at these prices, then the costs of the operation would be approximately zero. However, at a given point, the slope of an iso-value ellipse is equal to

$$\frac{-dq_S^0}{dq_L^0} = \frac{P_L - (1/\rho)[\sigma_{SL}q_S^0 + \sigma_{LL}q_L^0]}{P_S - (1/\rho)[\sigma_{SL}q_L^0 + \sigma_{SS}q_S^0]}$$

Thus, there will generally be costs incurred by the government when changing the composition of a given market value of the federal debt.

[18] For references and a discussion concerning the empirical evidence regarding this condition, see fn. 1.

[19] Equation (10) is derived by equating the slopes of equations (6) and (9) for given aggregate values of private and Treasury securities. While the left-hand side of (10) initially involves covariances (σ_{ES}, σ_{EL}), they drop out when simplifying further.

[20] It should be noted that in principle it is possible to solve the maximization problem explicitly in terms of the optimal quantities of short- and long-term Treasury securities (q_S^*, q_L^*), but the resultant expression is too difficult to interpret.

[21] Although it is unlikely that the government authorities ever operate on the portion of a given iso-value ellipse that is concave to the origin, this may be easily determined by the implementation of a market experiment. In particular, if a marginal increase in either short- or long-term Treasury securities reduces the aggregate value of Treasury securities, then the current combination of securities is on the concave portion of an iso-value ellipse.

In this case, the aggregate value of private securities is the lowest attainable value consistent with the given aggregate value of Treasury securities V_G^0. In moving from the concave to convex portions of the iso-value ellipse to pursue a stimulative policy, the government may inadvertently incur a cost even though the aggregate value of Treasury securities remains unchanged.[22] However, even if there is a cost, it may be stimulative to achieve a combination of securities on the convex portion of the iso-value ellipse and then recoup the cost by selling Treasury securities in combinations along the locus of tangencies of iso-value lines with iso-value ellipses.

An additional case which involves alternative covariance assumptions ($\sigma_{ES}, \sigma_{EL} < 0$, $\sigma_{SL} > 0$) is also of interest. As indicated previously, this case is consistent with a model presented by Tobin (1963). The implication of these covariance assumptions is that the most stimulative policy is determined by the tangency of V_E^0 with V_G^0 in Figure 1, where, in this case, $V_E^0 > V_E^*$. This follows directly from the assumption that $\sigma_{ES}, \sigma_{EL} < 0$, which implies from (6) that issuing Treasury securities is expansionary. It should be noted, however, that the interest costs of the Treasury have been neglected in formulating the objective of federal debt management policy.

B. *Restrictive Federal Debt Management Policies*

If the government authorities are concerned about inflationary pressures due to excess aggregate demand in the nonfinancial sector of the economy, then the optimal debt management policy may be one which implies an aggregate value of private securities V_E^r less than the maximum attainable value V_E^*. The range of possible restrictive policies encompasses $V_E^* > V_E^r > V_E^0$ which describes a set of parallel iso-value lines between V_E^* and V_E^0 in Figure 1. To implement the restrictive debt management policy involving a change from V_E^* to V_E^1 in Figure 1, for example, either of the operations involving the substitution of short- and long-term Treasury securities may be utilized. The resultant composition of Treasury securities actually selected may be determined by additional factors such as an interest-cost-minimization objective.

A neutral federal debt management policy, as advocated by Tilford Gaines and Milton Friedman, may also fall into the class of restrictive policies. A neutral debt management policy may be interpreted as one which maintains a constant composition of the Treasury debt held by the public through the jointly determined debt management policy of the government authorities. The proponents of this policy approach suggest that this method would alleviate the uncertainties involving the effects of discretionary debt management policies along with increasing the effectiveness of monetary policy. However, if the neutral debt management policy does not maintain the composition of the Treasury debt that is consistent with equation (10), then the maximum stimulative policy has not been achieved. Furthermore, it is most likely the case that a discretionary debt management policy should be implemented unless the government authorities cannot identify the distributions of any of the parameters describing the economy.

V. Summary of Conclusions

The relationship between Treasury and private securities investigated in this paper follows from an equilibrium model of financial asset markets incorporating the optimizing behavior of economic agents in a world characterized by uncertainty. The existence of this relationship depends on the existence of nonzero covariances between investors' assessments of the future prices of Treasury and private securities. In the case where the

[22] In Figure 1, for example, in moving directly from the tangency of V_E^0 with V_G^0 to the tangency of V_E^* with V_G^0, there is an unambiguous cost to the government as long as all prices are greater than zero. The possibility of being on the concave portion of an iso-value ellipse does not imply that the prices of short- and long-term Treasury securities are negative. Instead, it implies that for an increase in the quantity of long-term Treasury securities, for example, the relevant elasticicities and market values are such that $\eta_{SL}(P_S q_S^0/P_L q_L^0) + \eta_{LL} < -1$, where $\eta_{SL} = (\partial P_S/\partial q_L)(q_L^0/P_S)$, and $\eta_{LL} = (\partial P_L/\partial q_L)(q_L^0/P_L)$.

future prices of Treasury and private securities have positive covariances, it was shown that increasing the size of the aggregate market value of Treasury securities reduces the aggregate value of private securities, given that the composition of short- and long-term Treasury securities remains constant. In addition, to keep the aggregate value of private securities constant, it was demonstrated that short- and long-term Treasury securities may be substituted in a manner that is inversely proportional to the ratio of the covariances of their future prices with the future price of equity.

Alternative federal debt management policies were investigated in the context of the impact on the aggregate value of private securities for changes in the composition of a given aggregate value of Treasury securities. The aggregate value of private securities was assumed to be directly related to the level of economic activity through Tobin's q. Thus, in this framework, the debt management policy that achieves maximum economic stimulation is one which maximizes the aggregate value of private securities. The acceptance of this policy implies that the ratio of the market prices of short- and long-term Treasury securities minus their corresponding expected returns should equal the ratio of the covariances of their future prices with the future price of equity.

REFERENCES

Kenneth J. Arrow, *Aspects of the Theory of Risk-Bearing*, Helsinki 1965.

G. O. Bierwag and M. A. Grove, "A Model of the Structure of Prices of Marketable U.S. Treasury Securities," *J. Money, Credit, Banking*, Aug. 1971, *3*, 605–29.

O. J. Blanchard and M. K. Plantes, "A Note on Gross Substitutability of Financial Assets," *Econometrica*, Apr. 1977, *45*, 769–71.

O. H. Brownlee and I. O. Scott, "Utility, Liquidity, and Debt Management," *Econometrica*, July 1963, *31*, 349–62.

B. M. Friedman and V. V. Roley, "Investors' Portfolio Behavior Under Alternative Models of Long-Term Interest Rate Expectations: Unitary, Rational, or Autoregressive," *Econometrica*, forthcoming.

Milton Friedman, *A Program for Monetary Stability*, New York 1960.

Tilford Gaines, *Techniques of Treasury Debt Management*, New York 1962.

M. J. Hamburger and W. L. Silber, "Debt Management and Interest Rates: A Reexamination of the Evidence," *Manchester Sch. Econ. Soc. Stud.*, Dec. 1971, *39*, 261–66.

John Maynard Keynes, *A Treatise on Money*, London 1930.

———, *The General Theory of Employment, Interest and Money*, New York 1936.

Axel Leijonhufvud, *On Keynesian Economics and the Economics of Keynes*, New York 1968.

J. Lintner, "The Valuation of Risk Assets and the Selection of Risky Investments in Stock Portfolios and Capital Budgets," *Rev. Econ. Statist.*, Feb. 1965, *47*, 13–37.

———, "The Aggregation of Investors' Diverse Judgements and Preferences in Purely Competitive Security Markets," *J. Finan. Quant. Anal.*, Dec. 1969, *4*, 347–400.

———, "The Market Price of Risk, Size of Market and Investors' Risk Aversion," *Rev. Econ. Statist.*, Feb. 1970, *52*, 87–99.

H. Markowitz, "Portfolio Selection," *J. Finance*, Mar. 1952, *7*, 77–91.

F. Modigliani and R. Sutch, "Innovations in Interest Rate Policy," *Amer. Econ. Rev. Proc.*, May 1966, *56*, 178–97.

——— **and** ———, "Debt Management and the Term Structure of Interest Rates: An Empirical Analysis of Recent Experiences," *J. Polit. Econ.*, Aug. 1967, *75*, 569–89.

J. Mossin, "Security Pricing and Investment Criteria in Competitive Markets," *Amer. Econ. Rev.*, Dec. 1969, *59*, 749–56.

Richard A. Musgrave, *The Theory of Public Finance*, New York 1959.

W. D. Nordhaus and H. C. Wallich, "Alternatives for Debt Management," in *Issues in Federal Debt Management*, Fed. Reserve Bank Boston, *Monetary Conference Series*, no. 10, Boston 1973, 9–25.

A. M. Okun, "Monetary Policy, Debt Management, and Interest Rates: A Quantitative Appraisal," in *Fiscal and Debt Management Policies: Commission on Money and*

Credit, Englewood Cliffs 1963, 331–80.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

V. V. Roley, "A Structural Model of the U.S. Government Securities Market," unpublished doctoral dissertation, Harvard Univ. 1977.

E. R. Rolph, "Principles of Debt Management," *Amer. Econ. Rev.*, June 1957, *47*, 302–20.

R. H. Scott, "Liquidity and the Term Structure of Interest Rates," *Quart. J. Econ.*, Feb. 1965, *79*, 135–45.

W. F. Sharpe, "Capital Asset Prices: A Theory of Market Equilibrium Under Conditions of Risk," *J. Finance*, Sept. 1964, *19*, 425–42.

W. L. Smith, "Debt Management in the United States," Joint Economic Committee, U. S. Congress, *Study of Employment, Growth, and Price Levels*, Washington 1960, 1–154.

J. Tobin, "Liquidity Preference as Behavior Toward Risk," *Rev. Econ. Stud.*, Feb. 1958, *25*, 65–86.

———, "Money, Capital, and Other Stores of Value," *Amer. Econ. Rev. Proc.*, May 1961, *51*, 26–37.

———, "An Essay on the Principles of Debt Management," in *Fiscal and Debt Management Policies: Commission on Money and Credit*, Englewood Cliffs 1963, 143–218.

———, "A General Equilibrium Approach to Monetary Theory," *J. Money, Credit, Banking*, Feb. 1969, *1*, 15–29.

N. Wallace, "The Term Structure of Interest Rates and the Maturity Composition of the Federal Debt," *J. Finance Proc.*, May 1967, *12*, 301–12.

Board of Governors of the Federal Reserve System, *Fed. Res. Bull.*, Washington, various issues.

Disadvantageous Syndicates in Public Goods Economies

By ANDREW SCHOTTER*

The formation of syndicates, unions, or cartels has recently been shown to lead to consequences which are counter to our economic intuition. Specifically, it has been shown (see Robert Aumann; Andrew Postlewaite and Robert Rosenthal) that syndication in private goods economies may be disadvantageous to the syndicate members in the sense that all of their imputations in the core of the syndicated economy may be (agent-by-agent) worse or at most no better than any imputation they might receive in the core of the unsyndicated economy. Aumann has even shown examples where all syndicated core points are worse than all unsyndicated core points and in which it is actually disadvantageous to be a monopolist.

In this paper I concentrate on the disadvantages of syndicate formation in public goods economies. I present an example of a simple public goods economy in which syndication is extremely disadvantageous in the sense that the unique core imputation of the syndicated players in the syndicated economy is exactly equal to that Lindahl equilibrium imputation which is most disadvantageous to the syndicate. Put differently, if the players in the syndicate decided not to form a syndicate but rather to act individually and play the "game of perfect competition" (as Jean-Claude Milleron calls it) by truthfully reporting their preferences to an auctioneer who announces parametric prices and allocates cost shares, then their final imputation from such behavior could not be worse than any imputation they would receive by forming a syndicate and bargaining in unison.[1] This result is significant because syndicates, such as labor unions, etc., many times are formed strictly with an eye towards private goods consumption (i.e., salary and fringe benefits). Consequently, even if they are advantageous in that endeavor and do increase the syndicate members' allocation of private goods, their detrimental effects with respect to public goods may cause them, on balance, to be disadvantageous.

In order to analyze this subject intelligently, I will first discuss the difference between what is generally called a "coalition" and what is meant when we use the word "syndicate" in game theory. Then, before presenting the analysis, I will present some problems that exist in the definition of characteristic functions for public goods economies. Finally, an example will be presented which, in a public goods context, exemplifies disadvantageous syndication in the spirit of Aumann, and Postlewaite and Rosenthal. This will be followed by a simple intuitive explanation of exactly why syndication may be disadvantageous in public goods economies.

I. Coalitions and Syndicates

As Lloyd Shapley has pointed out, the concept of blocking in game theory is frequently misunderstood. While the word connotes "foul play" and disruption in its everyday use, in game theory it merely indicates that a set of players who form a coalition to block an imputation are together to get the most for themselves *using only their own resources*. As Lester Telser points out, the unrestricted formation of coalitions is the

*Associate professor, New York University. I would like to thank the Office of Naval Research whose partial support through Contract Number N0014-78-C-0598 made the research involved in this paper possible. In addition, I would like to thank George Borts and an anonymous referee for the many helpful comments and suggestions that helped me to revise this paper. All errors are of course mine alone.

[1]This, of course, implies that all other agents also play the game of perfect competition and truthfully report their preferences so that the true Lindahl equilibrium is reached.

essence of competition and not a sinister or unethical act.

Syndicate formation is quite different, however. In an n-person game, a syndicate is a set of players $S \subset N$ who join and decide to act in unison.[2] Consequently, no subset of S will join a coalition with any member not in S unless all of the players in S also join. If the players who join such a syndicate or union are all of one type (see A. Charnes and Stephen Littlechild; Terje Hansen and Jean Jaskold-Gabzewicz) then the syndicate formed is in essence a monopoly or cartel. We would expect that this would increase (or at least not decrease) their imputation in the associated game over what they would have gotten if they had either not formed a syndicate or acted "competitively." We will find that this need not be true for public goods economies.

II. What a Coalition can Achieve for Itself

In classical game theory the idea of what a coalition can achieve for itself as depicted by the characteristic function is easily defined. Basically, the value of a coalition is that amount of utility R (in the case of transferable utilities) or that set of utility vectors $V_S \subset E^S$ (in the case of nontransferable utilities) that a coalition can guarantee itself *no matter what the remaining players do*. In "orthogonal games," as Shapley and Martin Shubik (1973) call them, this is unambiguously defined. (Market games without externalities are one example of an orthogonal game.) When externalities or public goods are present, the question of what a coalition can achieve for itself is not so easily answered, because the payoff to the coalition is directly related to what the remaining players in the game do. Faced with this problem, Rosenthal pointed out that the classical definition of the characteristic function and consequently of blocking in economies characterized by externalities may not be intuitively appealing. He outlines four types of behavior that can be expected from a countercoalition S^* in a public goods economy (or economy containing externalities) when S forms a coalition. Following Rosenthal and Donald Richter we call these modes of behavior o type, I type, G type, and IG type.

In this paper I will employ the IG and G types of behavioral assumptions only. The G-type behavior has a very simple explanation: if a coalition S forms, it can expect its countercoaliton S^* to take actions which determine group-rational or Pareto optimal imputations for itself. In other words, it is assumed that S^*, having been abandoned by S, will merely do the best it can for itself under the circumstances and maximize its joint utility. Under the IG type of assumption, coalition S^* again organizes its activities in a Pareto optimal or group-rational manner. This time, however, we require that each player's final imputation be individually rational or at least as large as it would be if that player acted individually and accepted the value of the game to himself. Consequently, these assumptions assume a rationality on the part of the countercoalition which states basically that "if you are not with us we will maximize without you even if our group maximization will benefit you. We are not going to hurt ourselves just to hurt you."

There are other types of rationality assumptions that could be made, however, namely the o-type and I-type assumptions. Under the o-type behavior if coalition S forms, it can rely on S^* taking that action which is absolutely worst for S. This may result in a payoff to the members of S^* which is not individually rational, that is, which might reduce their imputation below what they can guarantee themselves by acting alone, but this threat cannot be ruled out. This is the conventional assumption and the one that Duncan Foley used to derive his results. It has led to the nonexistence of the usual limit theorems concerning cores and competitive (in this case Lindahl) equilibria.

Finally, in I-type behavior S^* will counter

[2]The first mention of a syndicate (although the name was not used) was done by John von Neumann and Oskar Morgenstern in a simple three-person trading model, pp. 568–69. More recently, Morgenstern and Gerhard Schwoediauer have reported results which show that the disadvantageousness of syndication will not occur if the von Neumann-Morgenstern solution concept is used as opposed to the core. Also, Michael Maschler has shown that such disadvantageous results will not occur if the bargaining-set solution is used. These findings lead one to question the appropriateness of the core concept

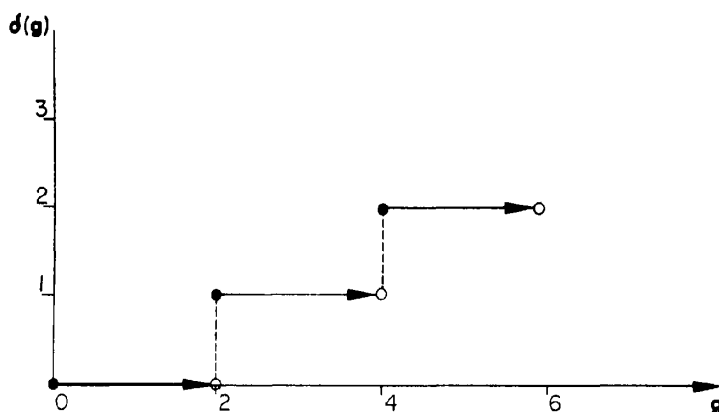



FIGURE 1 PUBLIC GOOD PRODUCTION FUNCTION $x_2 = \sigma(g)$

by organizing its activities to insure itself at least an individually rational payoff vector, that is, a vector that guarantees each player at least as much as they can guarantee themselves by acting alone.

The reason I do not employ either the o-type or the I-type rationality assumptions in my analysis is simple. The o-type behavior is purely spiteful behavior. Consequently it may require the countercoalition to "cut off its nose to spite its face" and such behavior may have disastrous consequences for coalition S^*. In addition, such an assumption was shown by Rosenthal to lead to unsatisfactory results. The I-type behavior, on the other hand, calls for a partial rationality which I feel is less desirable than the rationality described by either the G or the IG types of assumptions. In any case, whatever type of assumption the reader feels is preferable, this analysis will only concentrate on the G or IG type of assumption.

III

Using our rationality concepts, let us look at a simple example of a public goods economy to get an intuitive idea of why syndicate formation may be disadvantageous.

Let E be an economy with a set of three identical traders N indexed $i = 1, 2, 3$, all characterized by the following utility function:

$$U^i = (x_1^i)^{1/z} + bx_2 \tag{1}$$

where[3] $z > 1$, $b > 0$; x_1^i is the amount of private good x_1 that the ith individual consumes and x_2 is the total amount of a pure public good produced in the economy. Assume an initial endowment as follows:

$$w^1 = (x_1^1, g) = (1, 1) \tag{2}$$

$$w^2 = (x_1^2, g) = (1, 1) \tag{3}$$

$$w^3 = (x_1^3, g) = (1, 1) \tag{4}$$

where x_1^i is the private good in i's possession and g an all-purpose good[4] yielding no utility but which can be transformed into either the private good x_1 or the public good x_2 by the following production functions:

$$x_1 = \Psi(g) = (1/\gamma)(g) \quad \gamma > 0 \tag{5}$$

$$x_2 = \delta(g) = K/2 \tag{6}$$

where K is the closest even integer not greater than g. The function $\delta(g)$ states that the production function for x_2 is a step function with steps at all even integers. This is depicted in Figure 1.

From the model's description, it is clear that x_1 will never be traded and that the only reason to form coalitions in this economy would be to produce the public goods x_2. Let us assume that the parameters of the model

[3]The fact that the public good appears in each individual's utility function as a linear additive term is merely a convenience and not necessary for the example to work.

[4]The term g may be considered an endowment of labor in an economy where leisure has no utility.

(b, γ, and z) are such that the best a coalition of two can do for itself is to take its two units of g and transform them into one unit of x_2 via δ. This is equivalent to assuming that the sum of the marginal rate of substitution of x_1 for x_2 for the two players evaluated at $x_1^i = 1$, $x_2 = 0$ is strictly less than the marginal rate of transformation. Using the *IG* or *G* rationality assumption defined above we can try to specify a characteristic function that would describe this economy. However, such an attempt would fail since the function would not be superadditive.

To demonstrate why this is so, consider the following description of the utility vector or vectors that each coalition in our economy can achieve for itself assuming either our *IG* or *G* assumption, and in addition assuming that no transfers of the existing endowments of private goods (the x_1^i's) are allowed between agents:[5] First each individual agent can guarantee himself $(1 + 1/\gamma)^{1/z} + b$ since he can assume that the counter coalition (a coalition of two agents) is capable of producing one unit of the public good (since they have two units of g between them), and producing one unit is the group rational thing for them to do, by assumption. A coalition of two, say i and j, cannot rely on the countercoalition producing any of the public good. Even if the countercoalition wished to, it only has one unit of g. Consequently, the singleton countercoalition would produce no public good, and transform all of its g into $(1/\gamma)$ units of the private good. The coalition of two could then guarantee itself only the following single utility vector, $x = (1 + b, 1 + b)$, in which each contributes one unit of g to build the public good and is left with one unit of the private good and one unit of the public good to consume. Finally, the grand coalition faces the null set as a countercoalition. Since it obviously can not rely on it to provide any of the public good, it will have to produce it itself through contributions from the members of the economy. To finance the public good they would have to collect two units of g from the three of them. Consequently, the set Y of utility vectors of the form

[5]This is, of course, just a simplifying assumption.

$$(7)\quad Y = \{y \mid y = [(1 + \frac{1 - a_1}{\gamma})^{1/z} + b, (1 + \frac{1 - a_2}{\gamma})^{1/z} + b, (1 + \frac{1 - a_3}{\gamma})^{1/z} + b]\}$$

$$0 \le a_i \le 1, \sum_{i=1}^{3} a_i = 2$$

constitute the set of utility vectors achievable by the grand coalition. Written out formally, what has just been described appears as

$$(8)\quad V((i)) = (1 + 1/\gamma)^{1/z} + b \quad i = 1, 2, 3$$

$$(9)\quad V(i, j) = x = (1 + b, 1 + b)$$

$$(10)\quad V(123) = Y = \{y \mid y = [(1 + \frac{1 - a_1}{\gamma})^{1/z} + b), (1 + \frac{1 - a_2}{\gamma})^{1/z} + b), (1 + \frac{1 - a_3}{\gamma})^{1/z} + b),]\}$$

$$0 \le a_i \le 1, \sum_{i=1}^{3} a_i = 2$$

It is easy to see that this is not a characteristic function since, for example, $V(i \cup j) < V(i) + V(j)$. However, this does not present a problem since we are interested only in the Lindahl equilibrium of this economy which is not a game-theoretical concept and can be defined independently of the characteristic function. The imputations associated with it appear as

$$(11)\quad L = \{l \mid l = [(1 + \frac{1 - a_1}{\gamma})^{1/z} + b), ((1 + \frac{1 - a_2}{\gamma})^{1/z} + b), ((1 + \frac{1 - a_3}{\gamma})^{1/z} + b]\}$$

$$0 \le a_i \le 1, \sum_{i=1}^{3} a_i = 2$$

This set L of Lindahl imputations, has a very simple explanation. Each $l \in L$ is characterized by a different vector of contributions

(a_1, a_2, a_3) where a_i specifies the amount of good g player i is being asked to contribute towards the construction of the public good. Since the sum of these contributions is 2, exactly one unit of the public good will be constructed. The remaining units of g that each agent has after paying this requested contribution, namely $(1 - a_i)$, can be used to produce $(1 - a_i)/\gamma$ units of x_1, the private good, through the production function $\Psi(g)$. Any vector (a_1, a_2, a_3) such that $0 \leq a_i < 1$ and $\Sigma_{i=1}^{3} a_i = 2$, determines an imputation in L because, at the individualized prices defined by any such vector, each agent would maximize his utility by contributing his called-for share and accepting the resulting bundle of public and private goods. In other words, at the announced vector of contributions each agent in the economy would prefer to have the bundle of private and public goods defined by the associated vector in L than any other bundle he could afford to consume at those implicitly defined prices. The resulting allocations are Lindahl allocations.[6]

To investigate the effects of syndication on the economy, let agents 1 and 2 form a syndicate and call them *Syn*. If we now tried to specify a characteristic function for the game defined by this economy, we would find that it would indeed exist, although it would define an inessential game or a game in which there was no incentive for coalition formation.

[6]It is interesting to note that imputations associated with the Lindahl equilibrium are identical to the core of the game defined by the characteristic function which assumes α-type rationality. Its characteristic function appears as

$$V((i)) = \left(1 + \frac{1}{\gamma}\right)^{1/z}$$

$$V(ij) = x = (1 + b, 1 + b)$$

$$V(123) = Y = \left\{y \,|\, y = \left(\left(1 + \frac{1 - a_1}{\gamma}\right)^{1/z} + b\right), \left(1 + \frac{1 - a_2}{\gamma}\right)^{1/z} + b\right), \left(1 + \frac{1 - a_3}{\gamma}\right)^{1/z} + b\right)\right\}$$

$$0 \leq a_i \leq 1, \sum_{i=1}^{3} a_i = 2$$

The reason why a characteristic function exists here and has a nonempty core is because of the α-type assumption which makes larger coalitions beneficial.

To see this, again assume the *IG* or *G* assumptions and make the usual assumption that syndicate members are treated equally within the syndicate (i.e., they share equally their joint contribution of g to the construction of the public good). We can define the characteristic function for this game as follows:

$$V(3) = \left(1 + \frac{1}{\gamma}\right)^{1/z} + b \tag{12}$$

Thus player 3 can rely on *Syn* to build one unit of the public good and it uses all of its g to build x_1.

$$V(Syn) = x = (1 + b, 1 + b) \tag{13}$$

Thus the syndicate can not rely on any public good construction by the singleton player 3 and maximizes its joint utility by building one unit of the public good.

$$V(3, Syn) = X = \tag{14}$$

$$\{x \,|\, x = [(1 + \frac{1 - a_3}{\gamma})^{1/z} + b, (1 + \frac{1 - \frac{2 - a_3}{2}}{\gamma})^{1/z} + b, (1 + \frac{1 - \frac{2 - a_3}{2}}{\gamma})^{1/z} + b]\}$$

$$0 \leq a_3 \leq 1$$

Here, the grand coalition can achieve any utility vector $x \in X$ that is determined by joint contributions of g which sum to 2 in which the syndicate splits the contribution not made by player 3.

Since this game is inessential, the core is nonempty, is unique, and is represented by the vector,

$$x = \left(1 + b, 1 + b, \left(1 + \frac{1}{\gamma}\right)^{1/z} + b\right) \tag{15}$$

which occurs when the syndicate alone builds the public good and player 3 gets a total "free ride." This is obviously the only core imputation. Any imputation that required player 3 to contribute a positive amount could be blocked by that player acting alone. This unique core imputation, however, is at the extreme end of

the set of unsyndicated Lindahl equilibria when viewed from the point of view of the syndicate's members. In other words, syndication determines a core imputation in which $a_1 = a_2 = 1$, $a_3 = 0$, or one in which the syndicate finances the entire public good by themselves without any contribution from the unsyndicated player. This is worse than they would have done if they had not formed a syndicate and accepted the imputation associated with any announced Lindahl equilibria. Any other unsyndicated Lindahl equilibria is at least as good for them.

IV. Why Syndication may be Disadvantageous

The explanation of why syndicates do so poorly in my example is quite simple. Basically it is because syndicates create an indivisibility in the bargaining process that works to the detriment of the syndicate's members. The reason for this is as follows:

In our simple economy, under any of our rationality assumptions G or IG, there exists an "optimal" size of coalition which maximized the blocking power of the players in it by maximizing the free ride they receive from the rest of the economy. In my example this is a coalition of size one. Before syndication, each player had an equal opportunity to form such a coalition or at least had an equal threat to. However, once the syndicate actually formed, the members in it lost a very powerful threat since none of them could ever be part of a coalition of this size. They had transformed themselves into an indivisible player whose size prevented them from using a bargaining threat that each one had separately before syndication. At the very outset of bargaining they are "too big."

The fact that the members of the syndicate have lost some of their bargaining or blocking power should be obvious simply from the fact that the syndicated game has a nonempty core, while the unsyndicated game does not. This can be attributed strictly to the fact that the syndicated players are less able to block imputations in the syndicated game than they are in the unsyndicated game. This is so because neither of the syndicated players can form a singleton coalition after syndication, and given our rationality assumptions, singleton coalitions have the maximum blocking ability since they have threats that large syndicates do not have.

V. Conclusion

The conclusion of this paper is simple: In economies that contain both public and private goods, syndication is a two-edged sword. While it might be beneficial to its members in providing them with a greater amount of private goods (salary, fringe benefits, retirement programs), it may diminish their utility in the possession of public goods since their size prevents them from getting the free ride that smaller agents can achieve. In forming a syndicate then, these various costs and benefits must be weighed.

REFERENCES

R. Aumann, "Disadvantageous Monopolies," *J Econ. Theory*, Feb. 1973, *6*, 1–12.

A. Charnes and S. C. Littlechild, "On the Formation of Unions in n-Person Games," *J. Econ. Theory*, June 1975, *10*, 386–402.

D. Foley, "Lindahl's Solution and the Core of any Economy with Public Goods," *Econometrica*, Jan. 1970, *38*, 66–72.

T. Hansen and J. Gabzweicz, "Collusion of Factor Owners and the Distribution of Social Output," *J. Econ. Theory*, Feb. 1972, *4*, 1–18.

M. Maschler, "An Advantage of the Bargaining Set Over the Core," *J. Econ. Theory*, Oct. 1976, *13*, 184–92.

J.-C. Milleron, "Theory of Value with Public Goods: A Survey Article," *J. Econ. Theory*, Dec. 1972, *5*, 419–77.

O. Morgenstern and G. Schwoediauer, "Competition and Collusion in Bilateral Markets," *Z. Nationalökon.*, Dec. 1976, *36*, 217–47.

A. Postlewaite and R. Rosenthal, "Disadvantageous Syndicates," *J. Econ. Theory*, Nov. 1974, *9*, 324–26.

D. Richter, "The Core of a Public Goods Economy," *Int. Econ. Rev.*, Feb. 1974, *15*, 131–43.

R. Rosenthal, "External Economies and

Cores," *J. Econ. Theory*, June 1971, *3*, 182–88.

L. **Shapley**, "Let's Block 'Block'," P-4779, Rand Corp., 1972.

——— **and M. Shubik**, "On the Core of an Economic System with Externalities," *Amer. Econ. Rev.*, Sept. 1969, *59*, 678–84.

——— **and** ———, "Game Theory in Economics," ch. 6: "Characteristic Function, Core, and Stable Set," R-904-NSF/6, Rand Corp., July 1973.

D. Starrett, "A Note on Externalities and the Core," *Econometrica*, Jan. 1973, *41*, 179–83.

Lester Telser, *Competition, Collusion, and Game Theory*, New York 1972.

John von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton 1947.

The Performance of Multiperiod Managerial Incentive Schemes

By PETER MURRELL*

A number of recently published articles have focused on the design of incentive schemes for economic agents in planned economies. (See John Bonin; Liang-Shing Fan; Jeff Miller and James Thornton; Martin Weitzman.) The incentive schemes have been designed with two objectives in mind.[1] First, before the beginning of the plan period, an enterprise manager must be induced to supply accurate information to plan authorities about production possibilities. The information supplied will be in the form of a target value of output level. Secondly, at the end of the plan period, the enterprise manager must be induced to report accurately the achieved value of production, whatever target value was reported initially.

On the surface, the Soviet incentive scheme (analyzed by Miller and Thornton, and Weitzman) is different from the scheme analyzed by Bonin (itself a generalization of Fan's scheme). The Soviet scheme contains not only a self-imposed production target but also a production target imposed by the planners. However, in the single period case, to which all authors restrict themselves, the planner's target is exogenous and therefore does not affect any enterprise decisions. Thus, effectively the Soviet scheme and Bonin's scheme are very similar and, not surprisingly, give exactly equivalent single period results.

For the single period case with uncertainty in production levels, the two objectives of incentive design are satisfied by the incentive schemes of the above papers.[2] First: "A plan resulting from self-imposed targets can reflect any risk of underfulfillment desired by the planner" (Bonin, p. 685). Thus, the planner can manipulate at will the tautness of the enterprise plan. In particular, in Fan's scheme the reported target will be that one with a probability of fulfillment of one-half. Secondly, the accurate reporting of accomplishments is also obtained: "... given any self-imposed target, the manager will always report the highest possible level of performance, i.e., the realized value of [the planned variable]" (Bonin, p. 685).

The aforementioned studies have all limited their analysis to a single period framework. The reason may be that the authors are keen to ensure that the "dynamic incentive problem" or "ratchet effect" does not appear in their model (see Bonin, p. 687; Weitzman, p. 252). The dynamic incentive problem arises when planners use present performance as a basis for future target setting. With such target setting by planners, producers will tend to bias downward present reports of output achieved in order to leave themselves with an easier target in the future. However, if the results reported above are a reasonable representation of the properties of the incentive systems, then the dynamic incentive problem will greatly diminish in importance. The incentive schemes, encouraging honest target setting, will obviate the need for planners to set targets. Planners will be able to rely safely on the enterprises' self-imposed targets. In a sense, one may say that the incentive schemes are much more powerful than previous authors have claimed. The incentive schemes solve not only the single period incentive

*Assistant professor of economics, University of Maryland. I would like to thank Jeff Miller and an anonymous referee for helpful comments.

[1]In the following description I use the terminology appropriate to an environment where central planners are solely concerned with the output targets of enterprises. As many authors have pointed out, the incentive scheme can be applied in situations where variables other than output are the chief object. Indeed, the appropriate environment can be thought of as planning within a large organization rather than in a whole economy. I use the planning-production terminology solely for ease of discussion.

[2]Miller and Thornton extend the results to the case where managerial effort affects output

problem, but also the dynamic incentive problem.[3]

There is, however, a lacuna in the above argument. Previous results, having been restricted to single period analysis in order to justify ignoring the dynamic incentive problem, do not reflect on the behavior of the incentive schemes in a multiperiod framework. Thus one must extend the analysis of the incentive schemes to a multiperiod framework in which the planners do not impose targets.

There is one change in environment which is crucial to the move from a single period to a multiperiod analysis: the presence of inventories. In the single period case, there is no reason why producers should disguise their production levels because undeclared output has no value. In the multiperiod case, unreported output from one time period can be kept as producers' inventory and be reported in future time periods. Thus the producer will have to weigh the value of honest declaration of production against the value of keeping extra inventories. In turn the change in the nature of the production reporting decision will alter the way in which the production target is chosen. In the ensuing sections, I show that such considerations will affect the performance of the incentive schemes.

I. Incentives In a Multiperiod Framework

The model of production will be identical to that of Weitzman and of Bonin. Actual output at time t, Y_t, is subject to uncertainties which are represented by the probability density function $f_t(Y_t)$. For ease of notation it will be assumed that $f_t = f$ for all t. None of the results presented in this paper depend upon this assumption. The function f is known by the enterprise manager, but not by the planners. It will be assumed that f is continuously differentiable at all points, and that $f(Y_t) \equiv 0$ for $Y_t \in (-\infty, 0]$.

At the beginning of every plan period, the enterprise manager picks the target level Z_t, and communicates this target to the planners. The target level Z_t is used by the planners for plan construction purposes. In order to simplify the analysis, I assume that Z_t has no influence on Y_t. (Martin Loeb and Wesley Magat have examined the performance of the incentive scheme when such an influence is explicitly introduced.) At the end of the plan period, the enterprise manager knows the realized value of Y_t, and must report and deliver an amount of output X_t. The manager, in fixing the level of X_t, can reduce or increase the level of inventories K_t. Therefore, there are two constraints:

$$X_t \leqq Y_t + K_t, \text{ and } K_{t+1} = K_t + Y_t - X_t$$

Once K_{t+1} is known, Z_{t+1} can be chosen.

I will focus on one particular case of the incentive schemes: the scheme introduced by Fan. Fan's scheme is a special case in that the costs of overprediction and underprediction of target level are symmetric. Thus, in each time period the bonus is given by[4]

$$\alpha X_t - \epsilon |X_t - Z_t| \quad \alpha > \epsilon > 0$$

As the scheme is operating in a multiperiod framework, it must be assumed that the manager has a time horizon T. Inventory left over after time T, K_{T+1}, is assumed to be valued at $V_{T+1}(K_{T+1})$ at time T. I will make the assumption that $V_{T+1}(0) = 0$, and that V_{T+1} is continuously differentiable with $0 \leqq V'_{T+1}(K_{T+1}) \leqq \alpha + \epsilon$ and $V''_{T+1} \leqq 0$. The valuation of an increment of terminal capital is set at less than or equal to $\alpha + \epsilon$, because this term represents the maximum bonus payment which the enterprise manager can obtain from producing and delivering one unit of output.

Define E_t as the expectation operator over $f(Y_t), \ldots, f(Y_T)$ and E as the expectation

[3]Some authors also use the term ratchet effect to refer to a situation where performance in time t affects the parameters of the incentive system in time t + 1. Such an effect might arise if planners adjust the incentive system for time t + 1 on the basis of the manager's output level in time t. In this paper, it is assumed that the incentive system is not changed from one time period to the next

[4]In the more general case the value of ϵ when $X_T > Z_T$ is different from that when $X_T < Z_T$. In Fan's scheme, X_T and Z_T are interpreted as profit levels.

operator in any one time period. The manager's task in choosing Z_t is to maximize[5]

$$(1)\quad E_t\{\sum_{s=t}^{T}\delta^s[\alpha X_s - \epsilon|X_s - Z_s|] + \delta^T V_{T+1}(K_{T+1})\}$$

where $\delta \leq 1$ is the manager's rate of time preference. Let us call the maximum value of (1) $V_t(K_t)$. Also define

$$W_t(K_t, Z_t) = \max_{X_t, Z_{t+1}, X_{t+1}} E_t\{\sum_{s=t}^{T}\delta^s[\alpha X_s - \epsilon|X_s - Z_s|] + \delta^T V_{T+1}(K_{T+1})|Z_t, K_t\}$$

$$\text{Now } V_t(K_t) = \max E\{(\alpha X_t - \epsilon|X_t - Z_t|)\delta^t + E_{t+1}[\sum_{s=t+1}^{T}\delta^s(\alpha X_s - \epsilon|X_s - Z_s|) + \delta^T V_{T+1}(K_{T+1})]\}$$

$$(2)\quad = \max_{Z_t, X_t} E\{(\alpha X_t - \epsilon|X_t - Z_t|)\delta^t + V_{t+1}(K_{t+1})\}$$

$$(3)\quad = \max_{Z_t} W_t(K_t, Z_t)$$

The above equations are found by simple application of the optimality principle of dynamic programming and conditional expectation properties.

II. The Manager's Optimal Program

The approach in finding the optimal policy must be one of induction backwards from time T. Policy at time T will depend critically on the valuation of postterminal capital stock: $V_{T+1}(K_{T+1})$. It is much more important to focus on the features of policy which are independent of V_{T+1}. Therefore, a description of policy at time T is given in the Appendix, where an additional assumption is placed on V_{T+1}. It is assumed that V_{T+1} is such that it will not be optimal to have $Z_T = X_T = 0$. This assumption solely implies that valuation of terminal capital stock must be heuristically derived from the incentive system. Having $Z_T = X_T = 0$ implies that postterminal valuation of inventory is much greater than the valuation derived from the incentive system. Hence, it is only proper to rule out the case $Z_T = X_T = 0$.[6]

Given the foregoing assumptions on V_{T+1}, it is shown in the Appendix that

$$\frac{dV_T(K_T)}{dK_T} = V'_T(K_T) = \alpha\delta^T$$

One only needs the value of $V'_T(K_T)$ in order to describe policies in previous time periods. Thus, to some extent, optimal policies for t = 1, . . . , T − 1 are independent of the particular form of V_{T+1}. As V_{T+1} can be expected to vary between different managers, I will focus on the policies that one can expect to be common to all managers. Therefore, in this section, attention is restricted to time periods 1 to T − 1.

Optimal policies are of two distinct kinds, depending on the size of the manager's time preference. The following two theorems show the relationship between the type of optimal policy and the relative sizes of the three parameters α, δ, and ϵ. Proofs are contained in the Appendix.

THEOREM 1: *If* $\alpha\delta \geq \alpha - \epsilon$, *then for* t = 1, . . . , T − 1, (i) $V'_t(K_t) = \alpha\delta^t$ *and* (ii) *the optimal policy is given by* $X_t = K_t + Y_t$ *for* $K_t + Y_t \leq Z_t$, $X_t = Z_t$ *for* $K_t + Y_t \geq Z_t$, *and* Z_t *is fixed by*

$$Prob\{Y_t < Z_t - K_t\} = \frac{\alpha(1-\delta)}{\alpha(1-\delta) + \epsilon}$$

THEOREM 2: *If* $\alpha\delta \leq \alpha - \epsilon$, *then for* t = 1, . . . , T − 1, (i) $V'_t(K_t) = \alpha\delta^t$ *and* (ii) *optimum policy is such that* $X_t = K_t + Y_t$ *and* Z_t *is given by*

$$Prob\{Y_t < Z_t - K_t\} = 1/2$$

Thus, it is observed that when the

[5]The manager is assumed to be risk neutral.

[6]In fact it can be shown that if $Z_T = X_T = 0$, the optimal policy is very similar to the optimal policy delineated in Section II. If $Z_T = X_T = 0$, then one can find an S such that: $T \geq S \geq 2$, $Z_{T-1} = X_{T-1} = Z_{T-2} = X_{T-2} = \ldots = Z_{T-S} = X_{T-S} = 0$, and policies in t = 1, . . . , t − S − 1 are given by Theorems 1 and 2. Therefore, the assumption that $Z_T \neq 0 \neq X_T$ solely restricts S to be zero, which seems to be intuitively reasonable.

manager's rate of time discount is high ($\alpha\delta \leq \alpha - \epsilon$), the incentive scheme has the properties which have been observed in the single period case. In each time period, the achieved production level is reported correctly. No inventories are carried over from one time period to another. The target value is fixed such that the enterprise has a probability of one-half of exceeding it. Therefore, the enterprise manager gives an accurate representation of production possibilities to the planners.

The performance of the incentive scheme changes drastically for those managers with a low rate of time discount ($\delta < 1, \alpha\delta \gtrless \alpha - \epsilon$). The value of inventories is now greater than the value of declaring output achievements which are above the target level. Thus the manager will not report any output above the target level, but rather keep such output as inventory. The increase in inventory will cause an increase in the target level chosen for the next time period. However, the target will be conservative compared to production possibilities. The manager will choose a target which has probability of fulfillment of over one-half.

The disadvantages inherent in the incentive system are most clearly seen when one examines the policy undertaken by an enterprise manager who has no time discount ($\delta = 1$). Such a manager will undertake a riskless policy, setting a production target equal to inventory level and keeping all new output as inventory to be reported as output in the next time period. Effectively, the incentive scheme induces a manager to keep a high level of inventory and, one period later, report this inventory to the planning authorities. The target level given to the planners is not a target in the true sense, but rather a report of past accomplishments.

III. Conclusion

The foregoing analysis shows that the incentive scheme will not function in a multiperiod framework. The planner cannot in general induce managers to report a reasonable target, nor an honest level for production accomplishments. Hence, one may conclude that we are still far away from the design of an incentive scheme which will obviate the need for the imposition of targets by planners in a Soviet-type economy.

One may argue that, as the incentive system works well for managers with a discount rate such that $\alpha\delta < \alpha - \epsilon$, the planners can improve performance by reducing ϵ/α. However, since it is absolutely necessary that $\epsilon > 0$, the incentive scheme will always fail for some managers whose rate of time discount is low (δ close to 1). Thus, whatever the relative sizes of α and ϵ, there will always be the likelihood that some managers will not react to the scheme in the way in which the planners would desire.

A type of incentive scheme similar to the one discussed above has been introduced in the Soviet Union. It is fitting to end on the importance of the above results to the performance of the scheme in that country. Soviet policymakers have advocated setting ϵ at a level greater than $3\alpha/10$ (see Weitzman for details). Thus, the incentive scheme will not produce the desired results for those Soviet managers whose rate of time discount is such that $\delta > 7/10$. Even if the planning period is one-year long, one would think that $7/10$ is an unreasonably low value of δ. Thus, the scheme will fail for most managers. When it fails, the scheme would encourage conservative target setting and the buildup of excess inventories. Thus, the incentive scheme would only exacerbate trends which are already well-known in the Soviet Union.

Appendix

A result which is necessary for the proofs of Theorems 1 and 2 is that if $V_{T+1}(\cdot)$ satisfies the assumptions given in the text, then $V'_T(K_T) = \alpha\delta^T$. Thus this result is proven first.

X_T is chosen to maximize $\delta^T\{\alpha X_T - \epsilon|X_T - Z_T| + V_{T+1}(K_{T+1})|Z_T, Y_T\}$

Given the restrictions on V'_{T+1} an optimal policy is

If $K_T + Y_T \leq Z_T$, then $X_T = K_T + Y_T$

If $K_T + Y_T > Z_T$, then

$X_T = K_T + Y_T$ if $V'_{T+1}(0) \leq \alpha - \epsilon$

$X_T = Z_T$ if $V'_{T+1}(K_T + Y_T - Z_T) \geq \alpha - \epsilon$

$X_T = Y_T - M(K_T)$ otherwise,

where $M(K_T)$ is defined by $V'_{T+1}[K_T + M(K_T)] = \alpha - \epsilon$. Therefore

$$W_T(K_T, Z_T) = \delta^T \int_0^{Z_T - K_T} \{(K_T + Y_T)(\alpha + \epsilon) - \epsilon Z_T\} f(Y_T)dY_t + A\delta^T$$

where $A = \int_{Z_T - K_T}^{\infty} \{(\alpha - \epsilon)(K_T + Y_T) + \epsilon Z_T\} f(Y_T)dY_T$ if $V'_{T+1}(0) \leq \alpha - \epsilon$

or $= \int_{Z_T - K_T}^{\infty} \{\alpha Z_T + V_{T+1}(K_T + Y_T - Z_T)\} f(Y_T)dY_T$ if $V'_{T+1}(K_T + Y_T - Z_T) \geq \alpha - \epsilon$ for all Y_T

or $= \int_{Z_T - K_T}^{Z_T + M(K_T)} \{\alpha Z_T + V_{T+1}(K_T + Y_T - Z_T)\} f(Y_T)dY_T + \int_{Z_T + M(K_T)}^{\infty} \{(\alpha - \epsilon)[Y_T - M(K_T)] + \epsilon Z_T + V_{T+1}[K_T + M(K_T)]\} f(Y_T)dY_T$ otherwise.

Z_T is chosen to maximize $W_T(K_T, Z_T)$. Therefore,

$$V'_T(K_T) = \frac{\partial W_T(K_T, Z_T)}{\partial K_T}$$

when the latter term is evaluated at the maximizing value of Z_T (see Samuelson, p. 34). Therefore,

$$\text{(A1)} \quad V'_T(K_T) = \delta^T \int_0^{Z_T - K_T} (\alpha + \epsilon) \cdot f(Y_T)dY_T + B\delta^T$$

where $B = \int_{Z_T - K_T}^{\infty} (\alpha - \epsilon) f(Y_T)dY_T$ if $V'_{T+1}(0) \leq \alpha - \epsilon$

or $= \int_{Z_T - K_T}^{\infty} V'_{T+1}(K_T + Y_T - Z_T) \cdot f(Y_T)dY_T$ if $V'_{T+1}(K_T + Y_T - Z_T) \geq \alpha - \epsilon$ for all Y_T

or $= \int_{Z_T - K_T}^{Z_T + M(K_T)} V'_{T+1} \cdot (K_T + Y_T - Z_T) f(Y_T)dY_T + \int_{Z_T + M(K_T)}^{\infty} (\alpha - \epsilon) f(Y_T)dY_T$ otherwise.

Now

$$\text{(A2)} \quad \frac{\partial W_T(K_T, Z_T)}{\partial Z_T} = \delta^T \int_0^{Z_T - K_T} - \epsilon f(Y_T)dY_T + C\delta^T = 0$$

where $C = \int_{Z_T - K_T}^{\infty} \epsilon f(Y_T)dY_T$ if $V'_{T+1}(0) \leq \alpha - \epsilon$

or $= \int_{Z_T - K_T}^{\infty} \{\alpha - V'_{T+1}(K_T + Y_T - Z_T)\} f(Y_T)dY_T$ if $V'_{T+1}(K_T + Y_T - Z_T) \geq \alpha - \epsilon$ for all Y_T

or $= \int_{Z_T - K_T}^{Z_T + M(K_T)} \{\alpha - V'_{T+1} (K_T + Y_T - Z_T)\} f(Y_T)dY_T + \int_{Z_T + M(K_T)}^{\infty} \epsilon f(Y_T)\, dY_T$ otherwise

Now let us evaluate $V'_T(K_T)$ in each of the three cases.

Case 1: If $V'(0) \leq \alpha - \epsilon$, then from (A2)

$$\int_{Z_T - K_T}^{\infty} \epsilon f(Y_T)dY_T = \int_0^{Z_T - K_T} \epsilon f(Y_T)dY_T$$

Substituting in (A1), we have $V'_T(K_T) = \alpha\delta^T$.

Case 2: If $V'_{T+1}(K_T + Y_T - Z_T) \geq \alpha - \epsilon$ for all Y_T there are two subcases:

$$\text{(i)} \quad \int_0^{\infty} V'_{T+1}(Y_T) f(Y_T)dY_T \leq \alpha$$

in which case $Z_T \geq K_T$. Then

$$\int_{Z_T - K_T}^{\infty} \alpha f(Y_T)dY_T - \int_0^{Z_T - K_T} \epsilon f(Y_T)dY_T = \int_{Z_T - K_T}^{\infty} V'_{T+1}(K_T + Y_T - Z_T) f(Y_T)dY$$

[7] Which of the latter two formulas for B is applicable will depend on the value of K_T. Thus, when K_T is such that we are on the boundary between the two regions, one must understand by $V'_T(K_T)$ the left or right derivative which is applicable. Since $V'_T(K_T)$ is found not to depend on K_T, then the left and right derivatives will be the same in all cases

Thus, substituting into (A1), one obtains $V'_T(K_T) = \alpha\delta^T$.

(ii) If $\int_0^\infty V'_{T+1}(Y_T)f(Y_T)dY_T > \alpha$

then $Z_T < K_T$.

The cases when $Z_T = 0$ have been excluded by assumption. Thus, from (A2),

$$\int_0^\infty V'_{T+1}(K_T + Y_T - Z_T)f(Y_T)dY_T = \alpha$$

Substituting into (A1), one obtains $V'_T(K_T) = \alpha\delta^T$.

Case 3: There are two subcases when C in (A2) takes on the third possibility.

(i) If
$$\int_0^{M(K_T)} V'_{T+1}(Y_T)f(Y_T)dY_T \leq \int_{M(K_T)}^\infty \epsilon f(Y_T)dY_T + \int_0^{M(K_T)} \alpha f(Y_T)dY_T$$

then $Z_T \geq K_T$ and

$$\int_{Z_T - K_T}^{Z_T + M(K_T)} V'_{T+1}(K_T + Y_T - Z_T)f(Y_T)dY_T = \int_0^{Z_T - K_T} - \epsilon f(Y_T)dY_T + \int_{M(K_T)+Z_T}^\infty \epsilon f(Y_T)dY_T + \int_{Z_T - K_T}^{M(K_T)+Z_T} \alpha f(Y_T)dY_T$$

Substituting into (A1) gives $V'_T(K_T) = \alpha\delta^T$

(ii) If
$$\int_0^{M(K_T)} V'_{T+1}(Y_T)f(Y_T)dY_T > \int_{M(K_T)}^\infty \epsilon f(Y_T)\, dY_T + \int_0^{M(K_T)} \alpha f(Y_T)\, dY_T$$

then $Z_T < K_T$ and

$$\int_0^{Z_T + M(K_T)} \alpha f(Y_T)\, dY_T + \int_{Z_T + M(K_T)}^\infty \epsilon f(Y_T)\, dY_T = \int_0^{Z_T + M(K_T)} V'_{T+1}(K_T + Y_T - Z_T)f(Y_T)dY_T$$

Substituting into (A1) gives $V'_T(K_T) = \alpha\delta^T$.

Thus, in all cases $V'_T(K_T) = \alpha\delta^T$, which is the result necessary for the proofs of Theorems 1 and 2.

PROOF of Theorem 1:

The proof will be one of induction backwards from t = T − 1. It is known that $V'_T(K_T) = \alpha\delta^T$ therefore if one can prove the theorem is true at time t using this information, then the theorem must be true at all previous time periods. X_{T-1} is chosen to maximize $\{\delta^{T-1}(\alpha X_{T-1} - \epsilon|X_{T-1} - Z_{T-1}|) + V_T(K_T)\,|\,Y_{T-1}, Z_{T-1}\}$. Given that $V'_T(K_T) = \alpha\delta^T$ and that $\alpha\delta \geq \alpha - \epsilon$ then optimal policy is

$$X_{T-1} = K_{T-1} + Y_{T-1} \text{ for } K_{T-1} + Y_{T-1} \leq Z_{T-1}$$
$$X_{T-1} = Z_{T-1} \qquad \text{for } K_{T-1} + Y_{T-1} \geq Z_{T-1}$$

Hence, $W_{T-1}(K_{T-1}, Z_{T-1})$
$$= \int_0^{Z_{T-1}-K_{T-1}} \{[(\alpha + \epsilon)(K_{T-1} + Y_{T-1}) - \epsilon Z_{T-1}]\delta^{T-1} + V_T(0)\}f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^\infty \{\delta^{T-1}\alpha Z_{T-1} + V_T(K_{T-1} + Y_{T-1} - Z_{T-1})\}f(Y_{T-1})\, dY_{T-1}$$

In order to find the maximizing value of Z_{T-1} set

$$\frac{\partial W_{T-1}(K_{T-1}, Z_{T-1})}{\partial Z_{T-1}} = 0$$

(A3)
$$\frac{\partial W_{T-1}(K_{T-1}, Z_{T-1})}{\partial Z_{T-1}} = -\int_0^{Z_{T-1}-K_{T-1}} \delta^{T-1}\epsilon f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^\infty \{\alpha\delta^{T-1} - V'_T(K_{T-1} + Y_{T-1} - Z_{T-1})\}f(Y_{T-1})\, dY_{T-1}$$
$$= -\delta^{T-1}\int_0^{Z_{T-1}-K_{T-1}} \epsilon f(Y_{T-1})\, dY_{T-1} + \delta^{T-1}\int_{Z_{T-1}-K_{T-1}}^\infty \alpha(1-\delta)f(Y_{T-1})\, dY_{T-1} = 0$$

Since $Prob\,[Y_{T-1} < Z_{T-1} - K_{T-1}]$
$$= \int_0^{Z_{T-1}-K_{T-1}} f(Y_{T-1})\, dY_{T-1} = 1 - \int_{Z_{T-1}-K_{T-1}}^\infty f(Y_{T-1})\, dY_{T-1}$$

solving equation (A3) gives

$$Prob\,[Y_{T-1} < Z_{T-1} - K_{T-1}] = \frac{\alpha(1-\delta)}{\alpha(1-\delta) + \epsilon} \leq \frac{1}{2}$$

When $\delta = 1$ (no time discount),

$$Z_{T-1} = K_{T-1}$$

Because Z_{T-1} is picked to maximize $W_{T-1}(K_{T-1}, Z_{T-1})$,

$$V'_{T-1}(K_{T-1}) = \frac{\partial W_{T-1}(K_{T-1}, Z_{T-1})}{\partial K_{T-1}}$$

where the latter term is evaluated at the maximizing value of Z_{T-1} (see Samuelson, p. 34). Hence,

$$V'_{T-1}(K_{T-1}) = \delta^{T-1}\int_0^{Z_{T-1}-K_{T-1}} (\alpha+\epsilon) f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^{\infty} V'_T(K_{T-1} + Y_{T-1} - Z_{T-1}) f(Y_{T-1})\, dY_{T-1}$$

Substituting from equation (A3):

$$V'_{T-1}(K_{T-1}) = \delta^{T-1}\int_0^{Z_{T-1}-K_{T-1}} (\alpha+\epsilon) f(Y_{T-1})\, dY_{T-1} - \delta^{T-1}\int_0^{Z_{T-1}-K_{T-1}} \epsilon f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^{\infty} \delta^{T-1}\alpha f(Y_{T-1})\, dY_{T-1} = \alpha\delta^{T-1}$$

Thus, by induction the theorem is proved.

PROOF of Theorem 2:

Again an induction proof is used with exactly the same method as in the previous proof. Using $V'_T(K_T) = \alpha\delta^T$, one can prove that the theorem is true at $T-1$ and by induction true at all previous time periods. X_{T-1} is chosen to maximize $\{\delta^{T-1}[\alpha X_{T-1} - \epsilon|X_{T-1} - Z_{T-1}|] + V_T(K_T)|Y_{T-1}, Z_{T-1}\}$. Given that $V'_T(K_T) = \alpha\delta^T$ and $\alpha\delta \leq \alpha - \epsilon$ the optimal policy is: $X_{T-1} = K_{T-1} + Y_{T-1}$.

Hence, $W_{T-1}(K_{T-1}, Z_{T-1})$

$$= \int_0^{Z_{T-1}-K_{T-1}} \delta^{T-1}\{(K_{T-1} + Y_{T-1})(\alpha+\epsilon) - \epsilon Z_{T-1} + V_T(0)\} f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^{\infty} \delta^{T-1}\{(K_{T-1} + Y_{T-1})(\alpha-\epsilon) + \epsilon Z_{T-1} + V_T(0)\} f(Y_{T-1})\, dY_{T-1}$$

At the maximizing value of Z_{T-1}, it is required that:

$$\text{(A4)}\quad \frac{\partial W_{T-1}(K_{T-1}, Z_{T-1})}{\partial Z_{T-1}} = -\int_0^{Z_{T-1}-K_{T-1}} \delta^{T-1}\epsilon f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^{\infty} \epsilon\delta^{T-1} f(Y_{T-1})\, dY_{T-1} = 0$$

Thus, $Prob\ [Y_{T-1} < Z_{T-1} - K_{T-1}) = 1/2$, and

$$V'_{T-1}(K_{T-1}, Z_{T-1}) = \frac{\partial W_{T-1}(K_{T-1}, Z_{T-1})}{\partial K_{T-1}} = \int_0^{Z_{T-1}-K_{T-1}} \delta^{T-1}(\alpha+\epsilon) f(Y_{T-1})\, dY_{T-1} + \int_{Z_{T-1}-K_{T-1}}^{\infty} \delta^{T-1}(\alpha-\epsilon) f(Y_{T-1})\, dY_{T-1}$$

By substituting in equation (A4),

$$V'_{T-1}(K_{T-1}) = \alpha\delta^{T-1}$$

Hence, the theorem is proved.

REFERENCES

J. P. Bonin, "On the Design of Managerial Incentive Structures in a Decentralized Planning Environment," *Amer. Econ. Rev.*, Sept. 1976, *66*, 682–87.

L.-S. Fan, "On the Reward System," *Amer. Econ. Rev.*, Mar. 1975, *65*, 226-29.

M. Loeb and W. A. Magat, "Success Indicators in the Soviet Union: The Problem of Incentives and Efficient Allocations," *Amer Econ. Rev.*, Mar. 1978, *68*, 173–81.

J. Miller and J. Thornton, "Effort Uncertainty and the New Soviet Incentive System," *Southern Econ. J.*, Oct. 1978, *45*, 432-46.

Paul A. Samuelson, *Foundations of Economic Analysis*, New York 1965.

M. L. Weitzman, "The New Soviet Incentive Model," *Bell J. Econ.*, Spring 1976, *8*, 251–57.

Inside the Monetarist Black Box: Comment

By Francisco L. Lopes*

In two recent papers, one of them in this *Review*, Jerome Stein has developed a dynamic model through which "the controversy about the roles of monetary and fiscal policies can be reduced to an econometric debate about empirical magnitudes" (1974, p. 883). The controversial issue of whether a bond-financed increase in real government spending has any permanent effect on the equilibrium rate of growth of nominal income is asserted to be reducible to the question of whether a key reduced-form coefficient P_5 is positive or not.

I argue that Stein has failed to work out fully the dynamic implications of his model. It has led him to the incorrect proposition that a positive P_5 coefficient implies a permanent positive effect of fiscal policy on the equilibrium rate of growth of nominal income. This comment shows that the correct conclusion is that fiscal policy will, with the exception of a very special case where Stein's result holds, produce over the long run a rate of growth of nominal income that either attains an equilibrium value equal to the rate of growth of the nominal money supply or increases explosively.

I

For the sake of brevity, I restate Stein's results without proof.[1] After some inessential simplifications, the equilibrium rate of growth of nominal income (which I denote by η) was stated as[2]

$$(1) \qquad \eta = \mu + \frac{P_5}{P_3^*} D\theta + \frac{P_4}{P_3^*} DG$$

where the variables are the rate of monetary expansion μ, the time derivative of the ratio of government interest-bearing debt to the money supply $D\theta$, and the time derivative of real government purchases of goods and services per unit of capital DG. The coefficients P_3^*, P_4, and P_5 are derived from a price change equation:

$$(2) \qquad \pi = P(U, \pi^*, m, G, \theta)$$

which relates the rate of inflation Π to the unemployment rate U, the expected rate of inflation π^*, real balances per unit of capital m, real government expenditure per unit of capital G, and the ratio of government interest bearing debt to the money supply θ, as follows:

$$(3) \quad P_3^* = mP_3 = m\frac{\partial \pi}{\partial m}; \quad P_4 = \frac{\partial \pi}{\partial G}; \quad P_5 = \frac{\partial \pi}{\partial \theta}$$

While it is reasonable to assume positive P_3^* and P_4, the sign of P_5 cannot be established on theoretical grounds.

Let the fiscal policy consist of a one-shot increase in real government expenditure per unit of capital financed through the sale of bonds. Since the rate of growth of the money stock is kept constant, the ratio of interest-bearing debt to money must increase. Assume, following Stein, that it grows steadily, so that in the long run we have $DG = 0$ but $D\theta > 0$. Therefore, according to (1), if P_5 is positive the rise in real government expenditure per unit of capital will permanently raise the equilibrium rate of growth of nominal income above the rate of monetary expansion ($\eta > \mu$).

This result is incorrect on two accounts. For one, it assumes that P_3^* and P_5 are constant in the long run, notwithstanding the facts that the ratio of interest-bearing debt to money (θ) is rising and that real balances per unit of capital (m) must be falling as a result of the rate of growth of nominal income being

*Department of economics, Pontificia Universidade Catolica, Rio de Janeiro.

[1]Whenever notation differed in his two papers, I have adhered to the most recent one.

[2]This is analogous to equation 39 of the 1976 paper, with the difference of not assuming that DG equals zero

above the rate of monetary expansion, since by definition:

$$\frac{Dm}{m} = \mu - n - \pi = \mu - \eta \tag{4}$$

where Dm is the time derivative of real balances per unit of capital and n is the equilibrium rate of growth of output and capital. Hence, Stein's result holds only if the derivatives of P_3^* and P_5 with respect to θ and m are identically zero, which is clearly an unwarranted assumption.

If P_3^* and P_5 change over time as a consequence of $D\theta$ and Dm different from zero, there will be no equilibrium rate of growth of nominal income.[3] Assume, for example, that P_3 and P_5 are constant over time, so that P_3^* must be falling when the rate of growth of nominal income is above the rate of money expansion. By (1), the growth of nominal income will accelerate indefinitely as long as the ratio of interest-bearing debt to money keeps increasing ($D\theta > 0$).

II

Another unwarranted assumption in Stein's analysis is that a one-shot increase in real government expenditures per unit of capital will, under a constant rate of monetary expansion, make the ratio of interest-bearing debt to money increase indefinitely. We have seen that if this were true, the rate of growth of nominal income would in general never stabilize again after any pure fiscal impulse. Explicit consideration of the government budget constraint, however, shows that fiscal policy may also lead to a constant ratio of interest-bearing debt to money in the long run ($D\theta = 0$) and hence to an equilibrium rate of growth of nominal income equal to the rate of monetary expansion.

The government budget constraint requires that government expenditures GpK minus government revenue TpK equal the issuance of new interest-bearing bonds DZ plus the increase in high-powered money DM, i.e.,

$$G - T = (DZ + MD)/pK \tag{5}$$

where K is the capital stock. Following Stein, this can be written as

$$G - T = D\theta + (1 + \theta)\mu m \tag{6}$$

or, rearranging terms:

$$D\theta = \frac{G - T}{m} - (1 + \theta)\mu \tag{7}$$

Starting from an initial equilibrium position where $D\theta = 0$, a one-shot increase in real government expenditures per unit of capital (G) will, *ceteris paribus*, cause the ratio of interest-bearing debt to money to increase over time, that is $D\theta > 0$. This by itself will tend to reduce $D\theta$, acting through the second term in the right-hand side of (7). On the other hand, however, $D\theta > 0$ implies, as a consequence of (1) and (4), that real balances per unit of capital (m) are falling over time, which has the effect of increasing $D\theta$. Hence, we cannot determine a priori whether $D\theta$ will increase, stay constant, or decrease over time.[4] If, however, it does decrease, the economy will again reach an equilibrium when $D\theta = 0$, with the rate of growth of nominal income equal to the rate of monetary expansion.

In summary, equations (1), (4), and (7) define a dynamic system which as a rule has a single equilibrium solution with $D\theta = 0$. Therefore, the *equilibrium* rate of growth of nominal income is equal to the rate of monetary expansion. If, however, the system is

[3]With the exception of course of the very special case in which the rates of change of P_3^* and P_5 are equal so that the ratio P_5/P_3^* stays constant over time

[4]The condition for a decreasing $D\theta$ can be derived from constraining its second-order derivative $D^2\theta$ to be less than zero. Differentiating (7) we have $D^2\theta = -(G - T/m)(Dm/m) - \mu D\theta < 0$ which, after substituting Dm/m from (4), leads to $((G - T)/m)(\eta - \mu) < \mu D\theta$ But from (1), $\eta - \mu = (P_5/P_3^*) D\theta$, so that, assuming $D\theta > 0$, we have $((G - T)/m)(P_5/P_3^*) < \mu$ as the condition for a decrease over time of $D\theta$. Note that in Stein's "monetarist" case where $P_5 < 0$, this holds automatically, but it may also hold in his "fiscalist" case where $P_5 > 0$

unstable, a fiscal policy may lead to explosive growth of nominal income.[5]

[5]Stein's result would hold only in the very special case where the ratio P_5/P_3^* stays constant in spite of changing P_3^* and P_5, and $D\theta$ does not decrease over time.

REFERENCES

J. L. Stein, "Unemployment, Inflation, and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867–87.

———, "Inside the Monetarist Black Box," in his *Monetarism*, Amsterdam 1976.

Inside the Monetarist Black Box: Reply

By JEROME L. STEIN*

The issues that motivated my "Black Box" and the *American Economic Review* papers are clearly stated by Franco Modigliani and Albert Ando and by James Tobin and Willem Buiter. Their points of view will place my reply to Francisco Lopes in its proper perspective. Modigliani and Ando wrote:

> After many years of sharp and sometimes acrimonious controversy between monetarists and their sympathizers on the one hand, and the "nonmonetarists" on the other, over the effects of monetary and fiscal actions on aggregate income, some consensus seems finally emerging at least over what to disagree about: the main significant area of remaining disagreement seems to be over the *short- and medium-run* impact of *pure macro fiscal actions* on *aggregate money income*. The monetarists ... hold, in Friedman's words that it "is certain to be temporary and likely to be minor" (1972), while the nonmonetarists ... hold that it must be substantial—at least for a good many quarters. [p. 17]

Tobin and Buiter voiced a similar opinion:

> ... advocates of fiscal measures were looking for short-term effects on aggregate demand, without ... contemplating that the stock of money should remain constant while the stocks of other assets grow. When Walter Heller argued that the tax cut of 1964 would increase demand and reduce unemployment he was talking about what would happen in 1970 or 1980 if the tax cut were even then the only change from the pre-1964 monetary and fiscal policies. In this context it was no answer to say that years of accumulation of debt in exclusively nonmonetary form would be contractionary. [p. 275]

*Brown University.

The issues then concerned the trajectory of nominal income, resulting from a *pure fiscal policy*, over the *medium run*. I considered the medium run to be a span of time wherein the labor-capital ratio (x_t) and the debt-money ratio (θ) need not attain their steady-state values. It is illegitimate to require that the debt-money ratio θ attain its steady-state values and not simultaneously require that the capital intensity attain its steady-state value. The reason is that real capital changes at a faster rate than does the nominal federal government debt. For example, from 1950 to 1970 the stock of real fixed-business capital (net of straight line depreciation) grew at 3.9 percent per annum, but the net federal government debt in nominal terms only grew at a rate of 1.6 percent per annum. If the debt-money ratio is to be a state variable so must be the capital intensity. Since the nonmonetarists consider the effects of a pure fiscal policy in the *medium run*, I followed their orientation and omitted the dynamics of the capital intensity and the debt-money ratio in my analysis of what is inside the monetarist "Black Box." My dependent variables were just the unemployment rate U, the rate of price change π, the anticipated rate of price change π^*, real balances per unit of capital m, the nominal rate of interest ρ, and the real wage w.

My medium-run steady state required that the unemployment rate and inflation rate converge to constants (U_e and π_e, respectively) and that there be no unanticipated inflation or deflation. The ratios of capital to labor and of debt to money were allowed to change.

The question to be investigated was whether a pure fiscal policy could significantly affect the trajectory of the growth of nominal income. I concluded that: (a) the height of the trend line relating the logarithm of nominal income to time would certainly be affected, (b) but the slope of that trend line,

which is the growth of nominal income, will converge in the medium run to $\mu + (P_5/P_3^*) D\theta$ as described by equation (1).

$$(1) \qquad DY_N/Y_N \equiv \eta = \mu + (P_5/P_3^*)\, D\theta$$

where $DY_N/Y_N \equiv \eta$ is the growth of nominal income, μ is the rate of monetary expansion, θ is the debt-money ratio, and $D \equiv d/dt$. (See also my 1976 paper, p. 254.) I argued that nonmonetarists claim that P_5 is positive and that monetarists argue that P_5 is nonpositive.

I. Medium-Run Stability

Lopes questioned my argument that, when the unemployment and inflation rates converge to constants, the growth of nominal income will converge to $\mu + (P_5/P_3^*) D\theta$. My rate of inflation equation is

$$(2) \qquad \pi = P(U, \pi^*, m, G, \theta)$$

where π is the rate of inflation, U is the unemployment rate, π^* is the anticipated rate of inflation, m are real balances per unit of capital, G are real government purchases of goods per unit of capital, and θ is the debt-money ratio.

In the medium run, the rate of inflation will converge to

$$(3) \qquad \pi_e = \mu - n + \frac{P_5}{P_3^*} D\theta$$

where $\mu - n$ is the growth of the money supply per unit of capital and

$$(4) \quad P_3^* \equiv m\frac{\partial P}{\partial m} \equiv \frac{\text{change in rate of inflation}}{\text{percent change in real balances/capital}}$$

Lopes argued that if $P_5 D\theta$ is not zero, then real balances per unit of capital will diverge or go to zero, since

$$(5) \qquad \frac{Dm}{m} = \mu - n - \pi_e = -\frac{P_5}{P_3^*} D\theta$$

For example, if $P_5 D\theta$ were positive, then m would go to zero. Hence, he argued, my $P_3^* \equiv m(\partial\pi/\partial m)$ would go to zero, and my steady-state rate of inflation (equation (3) above) would diverge.

This criticism results from a misunderstanding, for which I am responsible, of the meaning of my P_3^*. I treated P_3^*, which is like an *elasticity* (see equation (4) above) as a constant; and $P_3 \equiv \partial\pi/\partial m$ is not treated as a constant. Lopes assumed that P_3 was treated as a constant, and this led him to claim that P_3^* goes to zero and π_e diverges in the fiscalist case. In my analysis of the fiscalist case, P_3^* is a constant so that the decline in m is completely offset by the rise in $\partial\pi/\partial m$.

Empirically, the assumption of a constant P_3^* or a constant P_3 may be true or false. Given my implicit assumption of a constant "elasticity" P_3^*, my medium-run system is dynamically stable as I showed in my paper. I apologize for the ambiguity and thank Lopes for the opportunity to clarify this point.

II. The Convergence of the Debt-Money Ratio

In Section II of his paper, Lopes considers my differential equation in the debt-money ratio θ,

$$(6) \qquad D\theta = \frac{G - T}{m} - (1 + \theta)\mu$$

where G = real government purchases per unit of capital, T = real net taxes per unit of capital, μ = rate of monetary expansion, m = real balances per unit of capital, θ = debt/money, $D = d/dt$. He asks whether θ will converge. This is indeed a profound question which has engaged the attention of Carl Christ, Karl Brunner and Alan Meltzer, Tobin and Buiter, Alan Blinder and Robert Solow (1973, 1974, 1976), and Ettore Infante and myself. It is the subject of our current research.

Equation (6) above (i.e., Lopes' equation (7)) cannot determine whether θ converges, because m is a state variable and real net taxes per unit of capital (i.e., taxes less transfers) may be endogenous. The convergence of the debt-money ratio cannot be determined without the simultaneous consideration of the convergence of the capital intensity.

Consider a steady state in a growing monetary economy[1] where the capital intensity, debt-money ratio, unemployment and inflation rates are constant. My Black Box medium-run steady state implies that the unemployment and inflation rates converge as a result of changes in fiscal or monetary policies. Insofar as the debt-money ratio is changing as a result of fiscal policy, the real rate of interest will change. Then, the capital intensity will not remain at its previous steady-state value. Once the capital intensity starts to change, the medium-run equilibrium referred to above will also change. This is why a constant capital intensity and rate of inflation require that the debt-money ratio be constant. Note, however, that the convergence of θ is a long-run (growth) problem, not the medium-run problem which is the subject of dispute between the monetarists and the nonmonetarists.

Two major questions are: (a) Can a system involving an endogenous capital stock be stable if government deficits or surpluses are financed by bonds, but where the money supply grows at a constant rate? (b) What will be the trajectories of the growth of output and nominal income resulting from a steep increase in real government purchases per unit of capital? These long-run questions, involving the trajectories of θ and the capital intensity, were ignored in my 1974 and 1976 papers concerning the monetarist controversy; but they are the subject of current research. Lopes discerned a very significant problem.

[1]It is a synthesis of my 1971 (ch. 5), 1974, and 1976 papers.

REFERENCES

A. Blinder and R. M. Solow, "Does Fiscal Policy Matter?," *J. Publ. Econ.*, Nov. 1973, *2*, 319–37.

——— **and** ———, "Analytical Foundations of Fiscal Policy," in *The Economics of Public Finance*, Brookings Institution, Washington 1974.

——— **and** ———, "Does Fiscal Policy Still Matter? A Reply," *J. Monet. Econ.*, Nov. 1976, *2*, 501–10.

K. Brunner and A. Meltzer, "An Aggregative Theory for a Closed Economy," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976, ch. 2.

C. Christ, "Some Dynamic Theory of Macroeconomic Policy," *J. Monet. Econ.*, Jan. 1978, *4*, 45–70.

E. Infante and J. L. Stein, "Does Fiscal Policy Matter?," *J. Monet. Econ.*, Nov. 1976, *2*, 473–500.

F. Modigliani and A. Ando, "Impacts of Fiscal Actions in Aggregate Income and the Monetarist Controversy: Theory and Evidence," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976, ch. 1.

Jerome L. Stein, *Money and Capacity Growth*, New York 1971.

———, "Unemployment, Inflation, and Monetarism," *Amer. Econ. Rev.*, Dec. 1974, *64*, 867–87.

———, "Inside the Monetarist Black Box," in his *Monetarism*, Amsterdam 1976, ch. 3.

J. Tobin and W. Buiter, "Long-run Effects of Fiscal and Monetary Policy on Aggregate Demand," in Jerome L. Stein, ed., *Monetarism*, Amsterdam 1976, ch. 4.

Welfare Measures in a Multimarket Framework

By RICHARD E. JUST AND DARRELL L. HUETH*

E. J. Mishan (1968) has demonstrated the well-known partial equilibrium result that the area behind a competitive supply curve conditioned on fixed-input prices (producer's surplus) measures returns or quasi rents on fixed-production factors. It is also a simple matter to show that the area behind a derived demand for inputs (conditioned on fixity of other input and output prices) measures returns or quasi rents on fixed-production factors of the production process using the input.[1] These results suggest that when one estimates supply and demand in an intermediate market and calculates the associated producer's and consumer's surplus under the stated conditions, then the welfare quantities exactly measure quasi rents for the two groups of firms involved in selling and buying in that market, respectively. That is, when all other prices facing the selling and buying firms are uninfluenced by their (group) actions, the ordinary surplus quantities do not include effects on other groups, for example, final consumers (of which there would be none if all other prices were truly unaffected because of elastic supplies, etc.).

In contrast to this extremely partial approach, a number of other authors (see, for example, Harry Johnson; Melvin Krauss and David Winch) have attempted to approach welfare surpluses from a general equilibrium standpoint where all other prices in the economy are allowed to vary. These works have culminated in the recent paper by James Anderson (1974) which shows that welfare changes can be determined by comparing the change in income arising from production (the area behind the general equilibrium supply function) with the income effect of price changes in consumption (which, as Anderson critically notes, has been referred to as consumer's surplus).

Each of these approaches has serious and well-known shortcomings in applied welfare economics. In reality, the imposition of a quota, tax, etc. in an intermediate market, such as for wheat or crude oil, may have a substantial effect on final consumption prices

*Professor of agricultural and resource economics, University of California-Berkeley, and professor of economics, Brigham Young University, and associate professor of resource economics, University of Rhode Island, respectively. We would like to thank Andrew Schmitz, John M. Currie, and especially James Anderson for helpful comments. Giannini Foundation Paper No. 512, Rhode Island Agricultural Experiment Station Paper No. 1681

[1]For example, consider a producer with variable inputs $x_1, \ldots, x_n$; output y, associated respective prices $\gamma_1, \ldots, \gamma_n, p$; and quasi-concave production function $f(x_1, \ldots, x_n)$ with

$$f(x_1, \ldots, x_{k-1}, 0, x_{k+1}, \ldots, x_n) = 0, \qquad k = 1, \ldots, n$$

Suppose that profit-maximizing levels of variable inputs and output are given by $x_k(\gamma, p)$, $k = 1, \ldots, n$, and $y(\gamma, p)$ where (γ, p) is considered a parametric vector for the competitive producer. Producer quasi rents are given by the restricted profit function,

$$\pi(\gamma, p) = py(\gamma, p) - \sum_{k=1}^{n} \gamma_k x_k(\gamma, p)$$

where, by the envelope theorem,

$$\frac{\partial \pi}{\partial \gamma_k} = -x_k(\gamma, p), \qquad k = 1, \ldots, n$$

Since $x_k(\gamma, p)$ may be interpreted as the derived demand for x_k when prices $\gamma_1, \ldots, \gamma_{k-1}, \gamma_{k+1}, \ldots, \gamma_n$ and p are considered fixed, the change in profits or quasi rents associated with an input price change from γ_k^0 to γ_k^* is

$$\Delta\pi_n = \int_{\gamma_k^0}^{\gamma_k^*} \frac{\partial \pi_n}{\partial \gamma_k} d\gamma_k$$

$$= \int_{\gamma_k^0}^{\gamma_k^*} - x_k^n(\gamma_1, \ldots, \gamma_{k-1}, \gamma_k, \gamma_{k+1}, \ldots, \gamma_n, p) d\gamma_k$$

and measures precisely the change in area behind the derived demand for x_k. In other words, *a producer's surplus as consumer of any one of his inputs is the same as his surplus as producer and supplier of his output in the fixed-price situation.*

and quantities, and conversely. Using partial methodology, these general equilibrium effects would be assumed away and thus ignored; in many interesting problems this practice is unacceptable. On the other hand, measurement with the general equilibrium approach "is far from our reach" (Anderson, 1974, p. 762) because of the intractability of practically estimating responses of all prices and quantities in an economy. Furthermore, general equilibrium measures only have "a satisfying and useful interpretation at the proper level of analysis, that of the economy" (Anderson, 1974, p. 761) which is at a too-aggregated level to answer many specific questions about proposed policy.

This paper is an attempt to fall somewhere between the above two cases in an effort to capture the essential generality of a class of important empirical problems while maintaining tractability. The framework considered involves a vertically structured competitive sector of an economy where each industry in the sector produces a single product using one major variable input produced within the sector, and a number of other variable inputs originating in other sectors of the economy. Any number of fixed-production factors may originate either from within or outside the sector. Assuming the sector is only one of many users of the other inputs, the prices in other sectors will be taken as fixed or uninfluenced by the sector in question. However, the actions of any individual industry in the sector may affect all other prices and quantities within the sector. In this context this paper examines the welfare significance of the classical triangles behind supply and demand as well as the possibilities of measuring both the direct and indirect effects of intervention.

Although the results may not be surprising for some practitioners of applied welfare economics (since on occasion some of the propositions of this paper have been used without proper examination), Daniel Wisecarver points out that considerable confusion has existed in the literature regarding these points. For this reason, Wisecarver, along with Richard Schmalensee (1971, 1976) and Anderson (1976), began to investigate the relationship of surplus measures in input markets with those in output markets; but the cases considered thus far deal only with long-run equilibrium or infinitely elastic input supply which disregards the producer side of the problem. While the results due to Anderson (1974), noted above, strengthen these propositions from the standpoint of aggregate welfare analysis (since they suggest that the consumer and producer sides of the problem can be studied independently), it becomes clear in this paper that interpretation of the usual surplus triangles with respect to individual market groups changes with market level. In point of fact, in the framework of this paper it is possible to examine disaggregated welfare impacts on each affected market group. Nevertheless, the results here are consistent with Anderson (1974) and show that the overall impact of introducing a distortion in some intermediate market is reflected by the sum of areas behind the general equilibrium supply and demand functions in that market.

I. Consumer's Surplus in an Intermediate Market

This section considers the welfare significance of the triangle behind demand and above price in an intermediate market of a vertically structured sector. For notational convenience, assume that the industries in a sector are competitive and can be ordered so that each industry n producing y_n and facing output price p_n, $n = 1, \ldots, N$, uses as variable factor inputs the product y_{n-1} produced at the preceding industry level in the sector plus some subset of inputs $x = (x_1, \ldots, x_m)'$ with corresponding price vector γ produced in the rest of the economy. The restricted profit function for the industry is, thus,

$$\pi_n(\gamma, p) = p_n y_n^n(\gamma, p) - p_{n-1} y_{n-1}^n(\gamma, p) - \gamma' x^n(\gamma, p)$$

where profit-maximizing levels of output and inputs at given prices are denoted by $y_n^n(\gamma, p)$, $n = 1, \ldots, N$, and $x^n(\gamma, p)$, respectively. Now suppose that prices in all industries in the sector are related through competition at the industry level so that, as price p_j is forcibly altered, the entire price vector changes

(monotonically) following $p(p_j)$. All inputs purchased from other industries, however, are available in elastic supply from the standpoint of the sector.

As pointed out by Mishan (1968), evaluation of the welfare impact of such a distortion in this case requires looking beyond the buyers and sellers in market j. Consider first the effects on any industry n in the sector where $j < n$. By the envelope theorem, one finds

$$\frac{\partial \pi_n}{\partial p_j} = y_n^n \frac{\partial p_n}{\partial p_j} - y_{n-1}^n \frac{\partial p_{n-1}}{\partial p_j}$$

Integration for a specific change from, say, p_j^0 to p_j^* implies

$$(1) \qquad \Delta \pi_n = \int_{p_j^0}^{p_j^*} \frac{\partial \pi_n}{\partial p_j} dp_j = \int_{p_j^0}^{p_j^*} y_n^n \frac{\partial p_n}{\partial p_j} dp_j - \int_{p_j^0}^{p_j^*} y_{n-1}^n \frac{\partial p_{n-1}}{\partial p_j} dp_j$$

where $\Delta \pi_n$ denotes the change in quasi rents for industry n.

To interpret (1) when $j < n$, note that the first right-hand term is the change in area behind demand and above price in market n, denoted by ΔC_n,

$$(2) \qquad \Delta C_n = -\int_{p_j^0}^{p_j^*} y_n^n \frac{\partial p_n}{\partial p_j} dp_j = -\int_{p_n(p_j^0)}^{p_n(p_j^*)} y_n^n \, dp_n$$

This is clear since, with $j < n$, integration in (1) is along equilibrium quantities in market n as the supply curve (influenced by p_j) is being shifted. It should be noted, however, that ΔC_n is not calculated with respect to the usual ordinary demand curve. It is, in fact, calculated according to the sector equilibrium demand curve (which is equivalent to a general equilibrium demand under the small-sector assumption of this paper) which accounts for adjustments in other industries in the sector.

To interpret the remaining right-hand term of (1) when $j < n$, note that integration is along equilibrium quantities in market $n-1$ as input supply (influenced or represented by p_j) is altered. Hence, the resulting integral represents the change in the area behind demand and above price in market $n-1$,

$$(3) \qquad \Delta C_{n-1} = -\int_{p_j^0}^{p_j^*} y_{n-1}^n \frac{\partial p_{n-1}}{\partial p_j} dp_j = -\int_{p_{n-1}(p_j^0)}^{p_{n-1}(p_j^*)} y_{n-1}^n \, dp_{n-1}$$

Again, one should bear in mind that the relevant demand curve for input y_{n-1} is a general equilibrium demand rather than an ordinary demand (in the same sense as above).

Using (2) and (3) in (1) implies that

$$(4) \qquad \Delta \pi_n = \Delta C_{n-1} - \Delta C_n, \quad n = j+1, \ldots, N$$

or, upon solving the difference equation for ΔC_j,

$$(5) \qquad \Delta C_j = \sum_{n=j+1}^{N} \Delta \pi_n + \Delta C_N$$

where ΔC_N represents the change in final consumer's surplus (associated with consumption of the sector's final product). Thus, the change in the "consumer's surplus" triangle in market j associated with an alteration in price p_j measures the sum of changes in final product consumer's surplus plus quasi rents for all industries (related by imperfectly elastic demands) involved in transforming the commodity traded in market j into final consumption form.

Of course, the welfare significance of $\Delta \pi_n$ is the same as in Mishan (1968). Also, if y_N is used as an input by some industry $N+1$ facing elastic demand, then ΔC_N merely measures the change in quasi rents for that industry; and one need search no further (on the demand side) for welfare effects of changing p_j. On the other hand, if market N is a final goods market, then the term ΔC_N may hold welfare significance in the context of final consumption in one of several senses. That is, if ΔC_N is the Marshallian surplus calculated from an ordinary final goods demand, then the results of Robert Willig can be used to determine the closeness of approximation to the proper Hicksian concept. Thus,

to the extent that his conditions apply in the final goods market, equation (5) holds welfare significance for market j measurements. Alternatively, one can interpret ΔC_N as a Hicksian surplus if the demand curve used in calculating ΔC_N and generating intermediate market demands is a compensated demand. In the latter case the measurement ΔC_j in (5) holds the proper Hicksian welfare significance without approximation.[2]

It is now clear that the Schmalensee (1976) and Anderson (1976) results hold only in a special case. Anderson's derivation, assuming long-run competitive equilibrium (in which case $\Delta \pi_N = 0$), shows that $\Delta C_{N-1} = \Delta C_N$ so that consumer's surplus can be measured in either the input or the output market. Schmalensee, on the other hand, assumed perfectly elastic input supply (i.e., $j = N - 1$) and found that $\Delta C_N + \Delta \pi_N = \Delta C_{N-1}$. Since this is the case where Mishan's (1968) results suggest that producer's surplus in market N measures quasi rents for industry N, equation (4) reduces to Schmalensee's result that the input market consumer's surplus is equivalent to the sum of producer's and consumer's surpluses in the output market.

II. Producer's Surplus in an Intermediate Market

To interpret ΔS_j, the change in the triangle area behind supply and below price in an intermediate market j, consider the effect of a similar change in p_j on market n where $j \geq n$. In this case, demand rather than supply in market n is affected by the change so that integration in (1) is along equilibrium quantities supplied as demand (or output price if $j = n$) is altered. Hence,

$$\Delta S_n = \int_{p_j^0}^{p_j^*} y_n^n \frac{\partial p_n}{\partial p_j} dp_j$$

$$= \int_{p_n(p_j^0)}^{p_n(p_j^*)} y_n^n \, dp_n$$

and similarly for ΔS_{n-1} which suggests that (1) can be rewritten as

$$\Delta \pi_n = \Delta S_n - \Delta S_{n-1} \qquad n = 1, \ldots, j \tag{6}$$

Solving the difference equation in (6), one finds

$$\Delta S_j = \Delta S_0 + \sum_{n=1}^{j} \Delta \pi_n \tag{7}$$

where ΔS_0 represents the change in initial resource supplier's surplus. Thus, the change in the "producer's surplus" triangle in market j associated with a change in p_j measures the sum of the change in initial resource supplier's surplus plus quasi rents for all industries (related by imperfectly elastic supplies) involved in transforming the raw resource into the commodity at market level j.[3]

Again, the welfare significance of $\Delta \pi_n$ is clear from Mishan (1968). The welfare significance of ΔS_0, on the other hand, is clear from Mishan (1959) in the case of a basic resource or from Mishan (1968) if there is an industry 0 facing perfectly elastic supply of *all* inputs. In the latter case, ΔS_0 simply measures the change in quasi rents for industry 0; and one need search no further (on the supply side) for welfare effects of changing p_j.

III. Market, Sector, and Economy-Wide Analysis

Summing the consumer's and producer's surplus measures in market j, one finds

$$\Delta C_j + \Delta S_j = \Delta C_N + \Delta S_0 + \sum_{n=1}^{N} \Delta \pi_n \tag{8}$$

Hence, where market 0 is a resource market and market N is a final goods market (so that the related chain of markets, $n = 0, \ldots, N$, comprises an economic sector), it is found that the sum of changes in producer's and consumer's surpluses in the distorted market actually measures the change in total sector welfare (where no intervening supplies or

[2] Note that matters are more complicated if some other final goods price changes as a result of altering price in market j. Only the Hicksian version has the proper path independence of the line integral used to evaluate the change in the expenditure function; but this is well understood following Eugene Silberberg.

[3] It is noted here that Anderson (1974) corresponds to the special case of (7) in which $\Delta S_j = \Delta S_0$ and $\Delta \pi_1 = \ldots = \Delta \pi_j = 0$ since he assumed pure competition in long-run equilibrium with constant returns-to-scale technology.

demands are perfectly elastic). Furthermore, since other sectors are unaffected by this change, the change in welfare for the economy as a whole is also obtained.

Thus, it turns out that the restrictive perfect elasticity assumptions which have been made in applied classical welfare economics are unnecessary. Classical welfare measures defined with respect to sector (or, equivalently, general) equilibrium curves have validity regardless, at least in the restrictive (small-sector) economic framework employed in this paper. But, more importantly, it is found that the classical triangles provide an overall rather than a partial picture of welfare. Hence, the failure of perfect elasticity assumptions in a vertical market structure has no serious consequences so long as one is interested in aggregate welfare rather than the welfare of a particular set of producers or consumers.

On the other hand, if disaggregation of welfare effects into impacts on individual market groups is needed (as in many policy analyses), several possibilities exist. For example, general equilibrium curves in adjacent markets can be used to compute the change in quasi rents for a given industry using the difference equations in (4) or (6). Alternatively, one can use ordinary supply or demand curves directly to compute the change in surplus or quasi rent for a given industry (using the results of Mishan, 1968, on the supply side or those of fn. 1 on the demand side). Yet another alternative is provided at least in *ex post* analysis by using value-added data directly (noting that the change in value-added is simply the change in quasi rents or restricted profits) without estimating other supply and demand relationships. Using any of these three methods to determine intermediate industry effects, the change in resource owner's welfare and final consumer's welfare can be estimated following (7) and (5), respectively.

One may note that the results obtained here differ somewhat from earlier assertions by Arnold Harberger and others; that is, that the general equilibrium measure of a change in welfare is approximately equal to the sum of changes in national income plus final consumer's surplus. Moreover, an exact relationship rather than an approximation is obtained here (in the case of Hicksian measurements). Indeed, one finds that the change in total welfare in the sector ($\Delta C_j + \Delta S_j$) is equal to the change in national income in the sector, $\Delta(p_N y_N)$, plus the change in final consumer's surplus (ΔC_N) only when initial resource supply is perfectly inelastic (in which case $S_0 = p_0 y_0$) since use of $\Delta\pi_j = \Delta(p_j y_j) - \Delta(p_{j-1} y_{j-1})$ in (8) thus implies

$$\Delta C_j + \Delta S_j = \Delta C_N + \Delta S_0 + \sum_{j=1}^{N} \Delta\pi_j$$
$$= \Delta C_N + \Delta(p_N y_N)$$

But when resource supply is not perfectly inelastic, as is the usual case with labor (for example, Mishan, 1959), the resource supplier's surplus is somewhat less than $p_0 y_0$ (supply is upward sloping) so that

$$\Delta S_0 + \sum_{j=1}^{N} \Delta\pi_j < \Delta(p_N y_N)$$

The amount by which the change in sector income overestimates ΔS_j is exactly the area under the resource supply curve.

IV. Empirical Implications and Considerations

The implications of the results in this paper for econometric welfare studies are several. First, it is clear that the definition of producer's surplus with imperfectly elastic input supply, etc. is not in terms of the usual competitive industry supply curve that is conditioned on fixed-input prices. Instead, ΔS_j is defined as the change in area behind the supply curve that takes account of general equilibrium price and quantity adjustments in markets $0, 1, \ldots, j-1$. By contrast, if one calculates the area behind the supply curve which holds input price fixed, then one obtains rents to producers in that market only.

Thus, suppose one estimates a supply equation,

$$p_j = b_0 + b_1 y_j + b_2 p_{j-1} \tag{9}$$

in an intermediate market j. It is clear from the results above that, if one calculates the

"surplus" measure S_j^* using b_1 as the slope and holds input price p_{j-1} fixed, then the quantity one obtains, $S_j^* - (1/2)\, b_1 y_j^2$, is a measure of quasi rents only to producers selling in market j. That is, $S_j^* \neq S_j$ but, rather, $S_j^* = \pi_j < S_j$. To estimate S_j, it is necessary to use the competition-induced relationship between p_j and p_{j-1}. Suppose, for simplicity, this relationship is given by

$$(10) \qquad p_j = a_0 + a_1 p_{j-1}$$

Then, substituting (10) in (9), one obtains

$$(11) \qquad p_j = \frac{a_1 b_0 - a_0 b_2}{a_1 - b_2} + \frac{a_1 b_1}{a_1 - b_2} y_j$$

as the supply curve which takes account of varying input prices. The surplus measure associated with (11) is given by

$$(12) \qquad S_j = \frac{a_1 b_1}{2(a_1 - b_2)} y_j^2$$

which is greater than S_j^* and measures benefits for all producers in markets $n = 0, 1, 2, \ldots, j$.

A problem which frequently arises in econometric work, however, is that input and output prices are highly correlated because of their theoretical relationship (which, indeed, leads to (10)). In practice, this condition usually prevents statistical identification of the parameters in (9) and thus leads to subsequent estimation of the equation,

$$(13) \qquad p_j = b_0^* + b_1^* y_j$$

where input price p_{j-1} is deleted. But, clearly, the interpretation of parameters and associated surpluses in (13) is different than in (9). Indeed, equation (13) is equivalent to (11); hence, an estimate of b_1^* in (13) is an estimate of $a_1 b_1/(a_1 - b_2)$ in (11), and the surplus measure associated with (13), $1/2\, b_1^* y_j^2$, is the same as S_j in (12).

It is also clear that symmetrical arguments can be made with respect to the area under estimated demand curves for intermediate goods. That is, replacing p_{j-1} by p_{j+1} and S_j by C_j, all of the derivation in (9)–(13) continues to hold.

The essence of this approach can be generalized to consider problems of econometric estimation and identification in the case where the relationships in (9)–(13) are changing because of varying prices (or determinants) imposed on the sector by the rest of the economy. That is, one can estimate a sector equilibrium supply curve in market j in the vertical sector case (encompassing endogenous adjustments of $p_0, \ldots, p_{j-1}, p_{j+1}, \ldots, p_N$) by regressing the quantity supplied y_j^s on market j price and all determinants γ_j^- affecting resource suppliers and industries $1, \ldots, j$,

$$(14) \qquad y_j^s = y_j^s(p_j, \gamma_j^-)$$

Similarly, the market j sector equilibrium demand could be estimated in the form,

$$(15) \qquad y_j^d = y_j^d(p_j, \gamma_j^+)$$

where y_j^d is quantity demanded and γ_j^+ is a vector of all determinants associated with the rest of the economy affecting industries $j + 1, \ldots, N$ and final consumer demand. If the small vertical sector assumption applies, then these sector equilibrium functions are indeed general equilibrium curves and can be used to estimate the overall welfare effects of a distortion.[4]

On the other hand, if one estimates (ordinary) supply and demand in the same intermediate market by considering prices in vertically related markets, that is, using a supply function of the form,

$$(16) \qquad y_j^s = \tilde{y}_j^s(p_j, p_{j-1}, \gamma)$$

and a demand function of the form,

$$(17) \qquad y_j^d = \tilde{y}_j^d(p_j, p_{j+1}, \gamma)$$

respectively, then one can obtain (assuming identifiability) partial equilibrium welfare measures, that is, measures of welfare for only the producers and consumers in the market in question by integrating only with respect to p_j as suggested above. Even when identification of (16) and (17) is not possible

[4]Note that econometric identification of the relations in (14) and (15) would be determined in the same manner as for simultaneous estimation of ordinary supply and demand except that the list of determinants is somewhat different.

because of multicollinearity, one can estimate price relationships,

$$p_{j-1} = p^s_{j-1}(p_j, \gamma) \tag{18}$$

$$p_{j+1} = p^d_{j+1}(p_j, \gamma) \tag{19}$$

which specify how prices in related markets change with respect to a distortion in market j for a given set of determinants outside the sector. If no other distortions exist in the sector, then satisfaction of the usual first- and second-order conditions for each industry in the sector implies that (18) and (19) are uniquely determined. Using (18) in (14) or (19) in (15), one can thus in principle solve for the ordinary functions in (16) or (17), respectively.[5] Hence, the ordinary surpluses pertaining only to industries j and $j + 1$, respectively, can be estimated. This approach can be expanded to determine quasi rents in other industries in the sector or, alternatively, the other approaches suggested above may be employed for this purpose.

V. Conclusions

This paper has studied welfare measures in a vertically structured sector. Given the usual conditions required for validity of consumer's surplus measures in the final goods market and producer's surplus measures in the initial resource market, the major results are as follows. The area behind a general equilibrium demand curve in an intermediate market does not measure benefits to buyers in that market alone, but rather measures the sum of rents to producers selling in all higher markets (assuming no intervening market has perfectly elastic demand) plus final consumer's surplus. Symmetrically, the area behind the general equilibrium supply curve in an intermediate market measures not only rents for producers selling in that market, but also rents for all producers selling in more basic markets (assuming no intervening market has perfectly elastic supply) plus initial resource supplier's surplus. Where some markets exhibit perfect elasticity of supply (demand), then producer's (consumer's) surplus measures rents for producers (consumers) in all lower (higher) markets related by supplies (demands) which are not perfectly elastic. To attribute welfare effects to those involved directly in the market in question versus those in the rest of the economy (sector), one needs merely to compare the areas behind ordinary supply and demand with the areas behind general equilibrium supply and demand functions. A simple and practical approach to studying the distribution of welfare effects over all other market groups in a sector is thus to estimate areas behind general equilibrium supply and demand curves in the market of interest (the sum of which would provide a measure of welfare in the sector). Welfare effects on direct market participants are then separated out by subtracting areas behind ordinary supply and demand curves. If appropriate data exist, the distributional aspects for other industries can also be studied by separating out the respective values-added.

Fairly well-defined vertical market chains seem to exist in many cases where a basic commodity sequentially reaches higher stages of refinement as it passes through the marketing channel, for example, petroleum, minerals, fisheries, and agriculture. It seems that the results of this paper are of practical applicability in the subset of related problems where contemplated policies may lead to changes in prices within the sector but do not cause noticeable price variation outside the sector.

[5]For example, if one estimates equations (10) and (13), then one can use $b_1^* = a_1 b_1/(a_1 - b_2)$ and $b_0^* = (a_1 b_0 - a_0 b_2)/(a_1 - b_2)$ and equation (9) to solve for b_0, b_1, and b_2

REFERENCES

J. E. Anderson, "A Note on Welfare Surpluses and Gains From Trade in General Equilibrium," *Amer. Econ. Rev.*, Sept. 1974, *64*, 758–62.

———, "The Social Cost of Input Distortions: A Comment and a Generalization," *Amer. Econ. Rev.*, Mar. 1976, *66*, 235–38.

A. C. Harberger, "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay," *J. Econ. Lit.*, Sept. 1971, *9*,

785–97.

H. G. Johnson, "The Cost of Protection and the Scientific Tariff," *J. Polit. Econ.*, Aug. 1960, *68*, 327–45.

M. Krauss and D. Winch, "Mishan on the Gains from Trade: Comment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 199–200.

E. J. Mishan, "Rent as a Measure of Welfare Change," *Amer. Econ. Rev.*, June 1959, *49*, 386–95.

———, "What Is Producer's Surplus?," *Amer. Econ. Rev.*, Dec. 1968, *58*, 1269–82.

R. Schmalensee, "Consumer's Surplus and Producer's Goods," *Amer. Econ. Rev.*, Sept. 1971, *61*, 682–87.

———, "Another Look at the Social Valuation of Input Price Changes," *Amer. Econ. Rev.*, Mar. 1976, *66*, 239–43.

E. Silberberg, "Duality and the Many Consumer's Surpluses," *Amer. Econ. Rev.*, Dec. 1972, *62*, 942–52.

R. D. Willig, "Consumer's Surpluses Without Apology," *Amer. Econ. Rev.*, Sept. 1976, *66*, 589–97.

D. Wisecarver, "The Social Costs of Input-Market Distortions," *Amer. Econ. Rev.*, June 1974, *64*, 359–72.

Inflationary Expectations and Momentary Equilibrium: Comment

By JOHN S. PETTENGILL*

In a paper in this *Review*, Lewis Johnson derived some striking results about the possible effects of a change in the expected rate of inflation using what he calls the portfolio adjustment model. In particular, he concluded that an increase in the expected rate of inflation could increase the real rate of interest and lower the actual rate of inflation. These results appear to come from a very strange specification of how consumers plan their monetary transactions.

According to Johnson, each individual has a target level of money balances $L(y, r)$, determined by the level of real income y and the nominal rate of interest r. Because of transactions costs, the individual will not instantaneously adjust his stock of money to this target level (as required by the portfolio balance theory). Johnson rather assumes that at each moment the individual economic agent observes how far his actual money stock m is from his desired money stock $L(y, r)$, and decides to adjust his money stock towards his target at a rate proportional to this difference (Johnson's equation (1)). Having decided the rate at which he plans to adjust his money stock, he observes how fast it is adjusting towards (or away) from his target as a result of exogenous factors (such as the (expected) rate of inflation). Then, and here is the crux of the issue, if he finds that (as a result, perhaps, of a high level of expected inflation) his money stock is adjusting towards his target level *faster* than planned, he will enter the money market and make transactions to *slow* his rate of convergence towards his target (Johnson's equation (2)). That is, he will make transactions in the opposite direction of the one he wants to go because the natural flow of events is pushing him towards his target faster than equation (1) says he ought to move.

I cannot think of a transactions cost model of noninstantaneous adjustment which would cause a person to engage in costly transactions to *slow* his adjustment towards his target. It might be possible to patch up the model by assuming that this paradoxical situation (where cash balances naturally adjust faster than the individual had planned to adjust them) never occurs in practice. But the underlying conception of consumer behavior that allows such a paradox still remains. Surely a rational consumer would use information about his current cash flow and the rate at which his real cash balances are appreciating or depreciating when planning how rapidly to adjust his real cash balances towards his target. By contrast, this paper requires the consumer to decide his cash balance strategy before he knows this information, and does not allow him to revise his plans when he does find out about it.

*University of Virginia

REFERENCES

L. Johnson, "Inflationary Expectations and Momentary Equilibrium," *Amer. Econ. Rev.*, June 1976, *66*, 395–400.

Inflationary Expectations and Momentary Equilibrium: Reply

By LEWIS JOHNSON*

In his comment on my paper, John Pettengill questions the behavior of economic agents implicit in my equation (2) which defines the flow excess demand for money. Specifically, he finds it paradoxical that an individual should "engage in costly transactions to *slow* his adjustment towards his target" (p. 955). The situation might arise, for example, when an individual wealth holder has real cash balances in excess of his desired stock. By my equation (2), this individual might nonetheless find himself in the market acquiring additional real balances if the inflation rate is sufficiently high (and even though he must pay transaction costs to do so).

While such behavior may seem puzzling at first glance, it is not at all peculiar to my model. In fact, an exactly analogous result obtains in the modern adjustment cost theories of investment, as may be found in the piece by Arthur Treadway (as just one example from this extensive literature). In these models one derives an optimal plan for accumulating capital over time in the face of increasing marginal adjustment costs. This plan normally involves convergence to a "target" capital stock with the property that a firm so fortunate as to begin with that capital stock would merely maintain it. The optimal plan can be linearly approximated by a partial adjustment rule which makes the *net* rate of capital accumulation proportional to the discrepancy between the target and actual capital stocks. The firm's flow demand for capital goods is, then, the sum of this proportion of the discrepancy and the expected depreciation of the current capital stock. But this is exactly analagous to my equation (2) with inflation depreciating the stock of real cash balances. The firm can be in a position of having capital in excess of its target level and still be in the market to purchase capital if the depreciation rate is high enough. It is not too difficult a task to modify a typical cost-of-adjustment model to provide a model of partial adjustment of real balance holdings from which equation (2) will emerge, just as its analogue does in the context of investment theory. See the paper by Eliot Bradford for one example of this sort of model.

In the context of these cost-of-adjustment models it is possible, I think, to find some insight into the economics of Pettengill's Paradox. Consider an economic agent whose initial position involves having real cash balances in excess of his target level, and who expects a positive rate of inflation. If he does not at some time enter the market to slow the reduction of his real balances, then his holdings will reach the target level in finite time (at t^*). From that moment forward he will have to purchase real balances at a rate equal to the inflation rate times his target level of real balances *for all time*. But alternatively he can at some time prior to t^* begin a series of purchases (vanishingly small) which will result in lower marginal transactions costs for all time. In these models, optimal paths turn out to be of this latter type, converging asymptotically to the target level.

*Federal Reserve Board.

REFERENCES

E. **Bradford,** "Inflationary Expectations and Momentary Equilibrium: Comment," unpublished paper, Claremont Grad. School 1977.

L. **Johnson,** "Inflationary Expectations and Momentary Equilibrium," *Amer. Econ Rev.*, June 1976, *66*, 395–400.

A. **Treadway,** "Adjustment Costs and Variable Inputs in the Theory of the Competitive Firm," *J. Econ. Theory*, Dec. 1970, *2*, 329–47.

Monopolistic Competition and Optimum Product Diversity: Comment

By JOHN S. PETTENGILL*

In their recent article in this *Review*, Avinash Dixit and Joseph Stiglitz (henceforth D-S) studied whether under monopolistic competition there will be too many or too few products. They conclude that if the demand for each individual product in the industry is of constant elasticity, then the monopolistic competition equilibrium will have the optimum number of products, given the constraint that each firm (product line) must at least break even. Then they go on to study the case of variable-elasticity product demand curves. They conclude that there is no necessary relation between the elasticity of the product demand curves and optimum product diversity. But they conclude that for an "important" family of utility functions, if the elasticity of demand falls as the scale of output in each firm rises, then monopolistic competition will lead to fewer than the optimal number of firms/products (by definition, each firm produces a single product).

This result is false. In fact, it is fairly easy to show that, within their model, if the elasticity of demand for an individual firm's product falls as the scale of output increases, then monopolistic competition will lead to too many products. Section I contains a simple derivation of this result. Section II shows where D-S went wrong in their analysis. Finally, in Section III it is argued that D-S's general framework is inappropriate for the study of monopolistic competition as it is normally thought of by most economists.

I

Following D-S, suppose we are considering an industry in which there are a large number of possible products, some number n of which are produced. Let x_i be the output of each product. Let the total cost of producing product i be $a + cx_i$, where a is the fixed cost and c the marginal cost. Each of the n products faces an identical demand curve $p(x_i)$, given that the prices of the other products remain fixed. Let $g(x_i)$ be the elasticity of this demand curve at output x_i,

$$g(x_i) = -\frac{1}{p'(x_i)}\frac{p(x_i)}{x_i} \tag{1}$$

It is an important feature of D-S's model that $g(x)$ is independent of the number of other products in the industry (and even of the prices of those products, as long as they are held fixed), vide their equation (32).

Since all firms have the same cost structure and face the same demand, we may assume that in equilibrium all firms produce the same quantity x_e of their product. Each firm will produce that quantity x_e which equates its marginal revenue with its marginal cost c. By the well-known formula giving marginal revenue as a proportion of price, it follows that

$$p(x_e)\left(1 - \frac{1}{g(x_e)}\right) = c \tag{2}$$

By definition of monopolistic competition, firms will enter until profits are zero

$$p(x_e) = (a + cx_e)/x_e \tag{3}$$

Suppose that in a monopolistic competition equilibrium there are n firms, each producing x_e. If we eliminated one product, and used the same resources elsewhere in the industry, we could increase output of each remaining product by

$$\frac{(a + cx_e)}{c(n-1)} = \beta \tag{4}$$

*Assistant professor of economics, University of Virginia.

The net effect on the consumer's surplus[1] will be

$$(5) \quad (n-1)\int_{x_e}^{x_e+\theta} p(s)\,ds - \int_0^{x_e} p(s)\,ds$$

If n is large (and D-S's equation (32) is only valid when n is large), this reduces to

$$(6) \quad \frac{a+cx_e}{c}\,p(x_e) - \int_0^{x_e} p(s)\,ds$$

That is, if n is large, the consumer values the small increase in his/her consumption of each other product at very close to the price $p(x_e)$ he/she now pays for the last marginal unit of each. Substituting (2) and (3) into (6) yields

$$(7) \quad \frac{p(x_e)x_e}{1-\dfrac{1}{g(x_e)}} - \int_0^{x_e} p(s)\,ds$$

as a measure of the gain (loss) from eliminating one industry. If (7) is positive, then monopolistic competition leads to too many firms; if (7) is negative, the monopolistic competition equilibrium is characterized by less than the optimal number of firms.

Let us now try the case of demand curves with constant elasticity g,

$$(8) \quad p_g(x) = Bx^{-1/g}$$

Doing the integration we find that

$$(9) \quad \int_0^{x_e} Bs^{-1/g}\,ds = \frac{Bx_e^{(1-1/g)}}{1-1/g}$$

and (7) is precisely 0. The monopolistic competition equilibrium is the optimal one (given the constraint that the firms in the sector break even and there are no subsidies from outside the industry). This is of course the result in the first section of D-S.

Suppose, now, that the individual demand functions $p(x_i)$ facing each firm were such that their elasticity $g(x_i)$ were a monotonically decreasing function of x_i. Given x_e construct a hypothetical demand curve $\hat{p}(x)$ to be that constant elasticity demand curve which has the properties:

1) It has the same elasticity $\hat{g}$ everywhere that $p(x)$ has at x_e(i.e., $\hat{g} = g(x_e)$).

2) It has the same price at x_e that p does (i.e., $\hat{p}(x_e) = p(x_e)$).

Since $p(x)$ has declining elasticity, $p(x)$ has a higher elasticity than $\hat{p}(x)$ in the region to the left of x_e (for $x < x_e$), and thus $p(x)$ lies below $\hat{p}(x)$ in this region. Formally,

$$(10) \quad p(x) < \hat{p}(x) \text{ for } x < x_e$$

It follows immediately from (10) that

$$(11) \quad \int_0^{x_e} p(s)\,ds < \int_0^{x_e} \hat{p}(s)\,ds = \frac{\hat{p}(x_e)x_e}{1-1/g} = \frac{p(x_e)x_e}{1-1/g(x_e)}$$

Consequently (7) is negative, and there are too many firms/products in the monopolistic competition equilibrium.

The above argument assumes that when a firm is eliminated total spending on the monopolistically competitive sector is constant. This is the assumption made by D-S in this section of their paper. When this assumption is relaxed one needs to add a correction term to (7). For example, if the elimination of a firm causes total expenditure to fall in the monopolistically competitive industry and rise in a competitive industry, consumer's plus producer's surplus is less than computed in (7), since marginal expenditures outside the industry are valued by consumers at the same rate as expenditures within the industry, but only expenditures within the industry generate producer's surplus.

II

Where, then, do D-S go wrong? They introduce a utility function of the form

$$(12) \quad U = x_0^{1-\gamma}\Big(\sum_{i=1}^{n} v(x_i)\Big)^{\gamma}$$

where x_0 is the quantity of output in the other

[1]The use of the consumer's surplus approach in this instance can be justified by noting that we are trying to ascertain whether welfare is increased or decreased by a move, not asking how much it is affected. Income effects can affect the magnitude of a change, but not its direction (since if there is no change in welfare, there is no bias from omitting the welfare effects).

industry. The elasticity of demand for the output of an individual firm is then given by

$$g(x) = -\frac{v'(x)}{xv''(x)} \tag{13}$$

Dixit and Stiglitz then define a function

$$\rho(x) = \frac{xv'(x)}{v(x)} \tag{14}$$

(which they call the elasticity of utility) and show (correctly) that if $\rho'(x)$ is positive, there are less than an optimal number of firms in the monopolistic competition equilibrium, given the break-even constraint.

Problems, however, arise when D-S introduce a specific form for $v(x)$,[2]

$$v(x) = (k + mx)^j, \quad m > 0,\ 0 < j < 1,\ k > 0 \tag{15}$$

Substituting (15) into (13), we find that

$$g(x) = \frac{(k/mx) + 1}{1 - j} \tag{16}$$

For all allowable values of m, j, and k, $g'(x) < 0$, and by Section I there will be too many firms under monopolistic competition. However if we compute (14) we have

$$\rho(x) = \frac{j}{(k/mx) + 1} \tag{17}$$

and $\rho'(x) > 0$, suggesting that there are too few firms at the monopolistic competition equilibrium.

The paradox is resolved when we realize that (15) is not a proper $v(x)$ function, because a proper $v(x)$ function must satisfy

$$v(0) = 0 \tag{18}$$

Otherwise the utility function (12) would have the property that one could increase the utility of each individual without limit by adding fictitious commodities to the list of products but producing zero quantities of them. Equation (18) is also implicitly required in the derivation of their result about the sign of $\rho'(x)$.

Suppose we redefine $v(x)$ to satisfy (18),

$$v(x) = (k + mx)^j - k^j \tag{15'}$$

This does not affect our computation of $g(x)$ in (16). But it does affect our computation of $\rho(x)$. It is easier to compute

$$\frac{x\rho'(x)}{\rho(x)} = k\left(1 - \frac{1 + jmx/k}{(1 + mx/k)^j}\right) - (k + mx)\left(1 - \frac{1}{(1 + mx/k)^j}\right) \tag{19}$$

(which of course has the same sign as $\rho'(x)$). The first term of (19) is always positive. As for the second term, if $j < 1$,

$$(1 + jmx/k) > (1 + mx/k)^j,\ x > 0 \tag{20}$$

(this can be shown by taking the binomial expansion of $(1 + mx/k)^j$, vide Tom Apostol, p. 44). It follows that the second term, and (18) itself, is always negative for $x > 0$. Dixit and Stiglitz' $\rho'(x)$ test now gives the same result as Section II's declining elasticity test, namely that monopolistic competition leads to too many firms.

III

Aside from these internal technical criticisms, I feel that D-S's framework is inappropriate to the analysis of monopolistic competition as most economists think of it. My first objection is that under their framework, each consumer consumes a small proportion of each product on the market. It is hard to think of any market where this is a plausible assumption. Some people do consume a large number of movies. But very few consumers see all the movies available. And seeing half of twenty movies is not preferable to seeing all of ten movies (which is required by the D-S framework).

My second objection is that one's normal

[2] Dixit and Stiglitz do not restrict $k > 0$. But if k were less than zero, one would have troubles with taking the fractional root of a negative number. If one tried to patch it up by defining $v(x)$ for negative $(k + mx)$ as the negative of the fractional root of the absolute value of $(k + mx)$, one introduces a nonconvexity into the utility function.

idea of monopolistic competition is that as the number of products in the industry increases, they become closer and closer substitutes (it becomes more difficult to differentiate additional products from the existing ones). Consequently, the elasticity of demand for each individual product should rise as the number n of firms/products increases, *ceteris paribus*. That is, the elasticity of demand for an individual product should be a function $g(x, n)$ of both output x and the number of products n.

Dixit and Stiglitz attempt to avoid explicitly including n as an argument to g by suggesting that elasticity of demand is a declining function $g(x)$ of the scale of output ($g'(x) < 0$). If the total size of the market is fixed, increasing the number of firms/products n decreases the output x of each, and thus increases the elasticity of demand for each individual product. But this is an artificial and unsatisfactory solution. Suppose that one doubled the number of firms/products and the number of consumers, so that the output of each product is unchanged. Surely this too should increase the elasticity of demand for each individual product. But it doesn't under the D-S specification. Conversely, it is possible (even plausible) that if a firm in a monopolistically competitive industry expanded its output it would face ever more severe competition, and its elasticity of demand would increase ($g'(x) > 0$). This should have nothing to do with whether if one increased the number of firms the elasticity of demand would increase. Yet the two are mutually incompatible in the D-S interpretation.

Whether $g(x)$ depends on n does not affect the analysis in Section I, given the assumption that n is large. But if the above reasoning is correct there is no necessary implication that $g'(x) < 0$ in monopolistically competitive industries, and thus no presumption (before examining an individual case) that monopolistic competition leads to too much or too little product diversity in this model. However, this conclusion (that under monopolistic competition there is no particular reason to expect too many or too few products) depends on the particular characteristics of the model. For example, Kelvin Lancaster, using a more realistic model in which the difference between products was measured by a quantitative index, concluded that monopolistic competition will lead to excessive product differentiation. And both models ignore the frequent accusation that advertising and other efforts to differentiate brands represent attempts to manipulate consumer's preference functions, and generate little or no direct utility.

REFERENCES

Tom Apostol, *Calculus*, Vol. 1, Waltham 1967.

A. Dixit and J. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *Amer. Econ. Rev.*, June 1977, *67*, 297-308.

K. Lancaster, "Socially Optimal Product Differentiation," *Amer. Econ. Rev.*, Sept 1975, *65*, 567-80.

Monopolistic Competition and Optimum Product Diversity: Reply

By Avinash K. Dixit and Joseph E. Stiglitz*

John Pettengill's comment on our 1977 paper has given us a welcome opportunity to give further thought to the problem. We find that his comment leads on to interesting issues, even though the substantive points he raises are invalid as stated.

I. Welfare Effects of Changing the Number of Firms

Pettengill tests whether there is an excessive number of firms in a monopolistically competitive equilibrium by a device of considerable expository merit He removes one firm, and redistributes the resources thus released equally over the remaining firms in the sector, to see if welfare can be improved. To do this correctly, we write n_e for the equilibrium number of firms and x_e for the output of each. With fixed cost a and constant average variable cost c, removing one firm releases $(a + cx_e)$ of resources, and this enables the output of each of the remaining $(n_e - 1)$ firms to be increased $(a + cx_e)/\{c(n_e - 1)\}$ The quantity x_0 of the numeraire good is unaffected by this, and the utility function (equation (31) of our paper) is

$$u = x_0^{1-\gamma}\{\sum_i v(x_i)\}^\gamma$$

Therefore welfare increases if and only if

$$(n_e - 1)v\left(x_e + \frac{a + cx_e}{c(n_e - 1)}\right) > n_e v(x_e)$$

This can be written

$$(n_e - 1)\left\{v\left(x_e + \frac{a + cx_e}{c(n_e - 1)}\right) - v(x_e)\right\} > v(x_e)$$

Using the first-order Taylor approximation for large n_e:

$$(n_e - 1)\frac{a + cx_e}{c(n_e - 1)}v'(x_e) > v(x_e)$$

Pettengill's equations (2) and (3) defining the equilibrium are

$$p_e\{1 - 1/g(x_e)\} = c$$
$$p_e = (a + cx_e)/x_e$$

where p_e is the equilibrium price and g is defined in equation (13) of Pettengill:

$$g(x) = -v'(x)/\{xv''(x)\}$$

Substituting, the criterion for welfare improvement is

$$\{x_e/(1 - 1/g(x_e))\}v'(x_e) > v(x_e)$$

or

$$\rho(x_e) > 1 - 1/g(x_e) \qquad (1)$$

where ρ is defined in our equation (35) $\rho(x) = xv'(x)/v(x)$.

Taking logarithmic derivatives of the definition of ρ,

$$\begin{aligned} x\rho'(x)/\rho(x) &= 1 + xv''(x)/v'(x) \\ &\quad - xv'(x)/v(x) \\ &= 1 - 1/g(x) - \rho(x) \end{aligned}$$

Therefore (1) holds if and only if $\rho'(x_e) < 0$, which is precisely the result contained in our equations (41) and (43).

Where, then, does Pettengill go wrong? It is clearly in making an invalid use of consumer's surplus. In fact $v(x_e)$ is not equal to

$$\int_0^{x_e} p\, dx$$

nor do the two stand in any fixed relation. If we maximize the utility function restated above, and evaluate the result in a symmetric case with zero pure profit, we find the relation

*Professors of economics, University of Warwick and Princeton University, respectively

$$p = \frac{1}{nx} \frac{\gamma\rho(x)}{\gamma\rho(x) + (1-\gamma)}$$

This can be found in our 1975 paper. Now we use the definition of $\rho(x)$ above to write this as $p = v'(x)\ h(x)$, where

$$h(x) = \gamma / \{nv(x)[\gamma\rho(x) + (1-\gamma)]\}$$

Integrating, we find that Pettengill's criterion bears no clear relation to $v(x)$, and is therefore not valid.

II. Presumptions Concerning Relation between g and ρ

Pettengill is quite right to realize that there are problems in using the family functions

$$v(x) = (k + mx)^j; \quad m > 0, 0 < j < 1, k > 0$$

but the problems are not what he thinks them to be. He thinks that since $v(0) > 0$, utility can be increased by introducing fictitious commodities with zero outputs. However, the original utility function was defined over all potential commodities, so they are all already there. But this tells us what the real problem is: for this family applied to a case with infinitely many potential commodities, the utility sum diverges. Matters might be rescued by defining

$$v(x) = \begin{cases} 0 & \text{for } x = 0 \\ (k + mx)^j & \text{for } x > 0 \end{cases}$$

This is a perfectly respectable increasing and concave function. It is discontinuous (and of course nondifferentiable) at $x = 0$, but that is immaterial since the demand functions for produced commodities are continuous. We only need its derivatives at x_e for the above argument. And here $\rho'(x) > 0$, i.e., the equilibrium has too few firms.

Pettengill's modification is

$$v(x) = (k + mx)^j - k^j$$

which has $\rho'(x) < 0$, i.e., the equilibirum has too many firms. This does weaken the "counterintuitive" presumption we claimed. But it does not suffice to establish the "intuitive" conclusion either. Therefore, albeit with somewhat greater diffidence, we continue to maintain that "the common view concerning excess capacity and excessive diversity in monopolistic competition is called into question" (1977, p. 304).

III. Which Model?

Pettengill concludes by expressing his view that our framework is inappropriate for the problem. His first point in this connection, that our approach must assume that each consumer consumes a small proportion of each product on the market, is untrue. As we clearly stated what is at issue is the convexity of Samuelsonian social indifference curves, and that can arise just as easily (and probably more commonly) because different consumers use different product types. Of course we need not rule out diversification by individual consumers: there are people who sample all thirty-one varieties of ice cream as well as those who stick to one favorite flavor.

Pettengill's preferred approach is the product characteristics model popularized by Kelvin Lancaster. For some purposes, particularly that of providing an intuitive feel for the kind of commodities that will be discriminated against in a market, that approach is extremely attractive. However, it suffers from the disadvantage that the derived demand functions are complex, and do not yield results in terms of parameters like the elasticity of demand that most economists have found intuitively helpful. By allowing such an understanding, the approach taken in our paper, and that of Michael Spence, should serve to complement the Lancaster approach.

The two approaches do differ in an important way, and provide different approximations that are valid in different situations. In most reasonably complex markets, the numbers of firms, consumers, products, and characteristics are all large but finite. Various idealized models take some of these to be infinite, or at least of a different order of magnitude compared to others. The case where the limit is competitive has attracted most attention. When the number of consumers and firms is allowed to increase, while

holding that of commodities fixed, there is increasing crowding in the commodity space (measured, for example, by cross elasticities), and the limit is evidently competitive. The Lancaster approach with a fixed finite characteristics space yields the same limit, since there must be increasing crowding in commodity space as the number of firms increases. These models may be good approximations in some cases. But there are several important instances where the observed market equilibrium is far from competitive, and a different idealization must be sought. This is the case with our approach, where the degree of crowding in commodity space does not increase, since the numbers of commodities and characteristics are of the same order of magnitude as the number of firms. We would not wish to claim that our approach is the uniformly best approximation, but we do not believe that such a claim can be made by the advocates of models which approximate to a competitive outcome either. Thus Pettengill's preferred approach must be seen as having restrictive validity, and the two approaches as complements rather than alternatives.

REFERENCES

A. K. Dixit and J. E. Stiglitz, "Monopolistic Competition and Optimum Product Diversity," *Amer. Econ. Rev.*, June 1977, *67*, 297–308.

______ **and** ______, "Monopolistic Competition and Optimum Product Diversity," econ. res. paper no. 64, Univ. Warwick, Feb. 1975.

K. Lancaster, "A New Approach to Consumer Theory," *J. Polit. Econ.*, Feb. 1966, *74*, 132–57.

J. Pettengill, "Monopolistic Competition and Optimal Product Diversity: Comment," *Amer. Econ. Rev.*, Dec. 1979, *69*, 957–60.

M. Spence, "Product Selection, Fixed Costs and Monopolistic Competition," *Rev. Econ. Stud.*, June 1976, *43*, 217–35.

On the Equivalence of Reference Price with Tariffs and Quotas

By ANTHONY Y.C. KOO*

The reported plan by President Carter (December 7, 1977, *New York Times*) to aid the steel industry combines financial assistance with a reference price for steel imports. The reference (sometimes called trigger) price system is a significant new development in the area of international commercial policy. To simplify the analysis, let us assume that the cost function of the domestic steel producer is given, and study how the introduction of reference price effects domestic price, output, and import level. We shall ignore the case in which the reference price is given and cost changes occur as a result of federal assistance.

Under the system, the reference (minimum) price for steel products will be based on the production costs of the most efficient foreign (in this case, Japanese) producer.[1] Any imports entering the United States and selling below the reference price will automatically be subject to a penalty of sharply higher "dumping" duties.

The question which immediately arises then is: In what way does the reference price system affect domestic output, price, and imports? More specifically, is it equivalent to, or different from, tariffs and quotas in their economic implications? It has been accepted that equivalence will generally hold between tariffs and quotas under the assumption of universality of competitiveness, that is, competitive domestic production, supply of imports and holding of quotas (see Jagdish Bhagwati, 1965, 1968, 1969; Hirofumi Shibata; Gopal Yadav). The *U.S.* steel industry hardly fits the description of a competitive industry. Thus the equivalence between quotas and tariffs will break down when measured in terms of the discrepancy between foreign and domestic prices (implicit tariff) (see Bhagwati, 1965) and/or of reducing imports (see Shibata) We shall assume for our present purpose that steel is competitively supplied from abroad and that the domestic steel industry operates as if it were a monopoly.

My analytic model consists of seven equations similar to those of Bhagwati (1969, ch 9). Since the result depends on discontinuities of functions which can be more easily handled by graphic treatment than by mathematics, the underlying equations are omitted here for lack of space. In Figure 1 it is assumed that a domestic monopoly is faced with an industry demand curve DD' and that the tariff rate shifts the foreign supply (reaching *U.S.* shores) schedule S_F upwards. The net demand schedule for the domestic monopolist is VMU, with marginal revenue schedule VR. Equilibrium is where the marginal revenue schedule cuts the marginal cost schedule, so that the monopolist's production (S_D) is at $0Q_0$, the domestic price at $0A$, the foreign price at $0W$, and import level (S_F) at MN.

*Michigan State University. My colleagues B T Allen, K. D Boyer, I. Chan, D S Hamermesh, and L. H. Officer have given me the benefit of their comments. I am also grateful to the anonymous referee who made useful suggestions including the example of discontinuous tariff and to George Borts for a suggestion to improve the diagrams.

[1]A brief reference to other similar but not directly related concepts is in order. The European Economic Community (*EEC*) countries have a reference price system with respect to import of fruits and vegetables In this case, unlike that of steel, price is determined by the most efficient producing region within the *EEC* rather than the rest of the world. Also similar to the reference price is the threshold price on grains established to control their imports into the *EEC*. There is a minimum (maximum) price at which the *EEC* countries buy (sell from the stocks) grains to maintain the floor (ceiling) of the grain support price. The threshold price is determined by deducting the cost of shipping grain overland to the population centers from the minimum price, that is, the threshold price is the minimum import price that the *EEC* will tolerate. Accordingly a variable levy, amounting to the difference between the c.i.f. import price and the threshold price, will be levied and will vary with any change in the difference between these two prices.

FIGURE 1

We then set a low reference price without tariff $0W'$ (or $0A'$ with tariff). The monopolist is confronted with a demand schedule of $VMM'N'D'$ (with $M'N'$ equal to the horizontal distance between A' and S_F with tariff). The corresponding marginal revenue curve of the monopolist is $VV'M'N'R'$.[2] However, the intersection point of the new marginal revenue schedule and the marginal cost schedule is the same as where the domestic production is at $0Q_0$, domestic price at $0A$ and import level at MN, as if there was no reference price but only tariffs in existence.

In Figure 2 we reproduce all the schedules of Figure 1 except that the reference price (again denoted by $0W'$ without tariff or $0A'$ with tariff) is higher. The demand schedule facing the monopolist is $VMM'N'U$ with corresponding marginal revenue curve $VHM'N'R'R$. This marginal revenue curve is cut by the marginal cost curve at E, I, and J, respectively. In this situation there are two local profit maxima, E and J.

To determine the point at which the maximum maximorum occurs, we compare the triangular areas EHI and $IM'J$. The area EHI represents an increase in profit (in this case a decrease because MC will be in excess of MR); similarly for the area $IM'J$. Therefore by moving the output which was determined at E to J, it is necessary that $IM'J$ be greater than EHI.

Consequently, the monopolist's supply (production S_D) is at $0Q_1$ where Q_1 is directly below J, and the domestic price is $0A'$ which is below $0A$ under tariff. But the import level will be reduced to JN' which is smaller than MN. In such a case the domestic production is larger while the domestic price is lower. The fact that such a case can be found may be a pleasant surprise to proponents of the Carter plan.

[2]The marginal revenue curve should be extended beyond R' to a point where it intersects the marginal revenue curve corresponding to DD' (not shown) directly under the point U on the DD' curve. But lack of space in the diagram does not permit this. Strictly speaking, the marginal revenue curve $VV'M'N'R'$ is discontinuous and consists of the segments VV' and $M'N'$. Where the intent is clear, the language in this paper is at times loose for the sake of convenience.

FIGURF 2

The result relies on two factors: First, the *U.S.* producers act as a monopoly; and second, that the reference price is set below the current market price. Taken together, what makes the reference price distinct and uniquely apart from the tariff and quota is that as a fixed price, it makes the marginal revenue curve of the net demand function discontinuous (with a horizontal portion). Once these factors are identified, the same discontinuity idea can be used to analyze other forms of commercial policy. Hence our analytic approach could be integrated with the previous literature on the equivalence of tariffs and quotas. For example, consider a discontinuous tariff as in Figure 3. Here S_F is the foreign supply curve without tariff. Suppose that the tariff is on imports only when the price P is less than P_0. The tariff schedule is $t(P_0 - P)$ for $P_0 > P$ as represented by ABS_F. Net domestic demand shifts from $VU'D'$ to $VCU''D'$, marginal revenue now has a new segment EF corresponding to $VCU''D'$, and MC intersects EF at J. Price accordingly falls from P_0 to P_1. Domestic output rises from Q_0 to Q_1.

What are the welfare implications of the reference price system in relation to tariffs and quotas? Given the domestic cost function, if the reference price level is low enough, the tariffs are in fact effective, and there will be no additional reduction of imports into the United States other than what is caused by the tariff barrier. If the reference price is set high enough relative to the domestic cost of

FIGURE 3

production, imports into the United States will reduce. The high reference price system appears to take on the characteristics of quota insofar as import reduction is concerned. But the resemblance is more apparent than real, because quota is generally determined by such considerations as avoidance of market disruption in the importing country rather than the cost of production of the most efficient foreign producer.

As time goes on, the most efficient producer may become even more efficient, or today's most efficient producer (in Japan) could be replaced by another producer in a different country (South Korea) where labor costs are even lower than those of Japan. Then the rules of the game call for a lowering of the reference price. Given the cost function of the domestic producer, a lowering of the reference price will result in an increase in the imports into the United States as shown in Figure 1 where the reference price is low relative to the cost function, and only tariffs are effective. Similarly, given the reference price if the (marginal) cost function of domestic steel producers should become higher, then according to my analysis more imports will occur. Graphically, this could be demonstrated by drawing a new MC' curve in Figure 2 (not shown) slightly higher than MC. The new MC' would intersect $M'N'$ to the left of J (not shown), resulting in larger imports. Thus under the reference price system, there is incentive for both foreign and domestic producers to increase efficiency. For the foreign producers, a lowering of cost means more steel can be exported to the United States. For domestic producers, they will face larger imports if they do not contain their cost function in relation to the reference price. Such a beneficial effect on the domestic price would not occur under the quota regime. In this respect, given a choice between the two, the reference price system should be considered a more welcome alternative than quotas.

REFERENCES

Jagdish Bhagwati, "On the Equivalence of Tariffs and Quotas," in Richard E. Caves, et al., eds., *Trade, Growth and the Balance of Payments—Essays in Honor of Gottfried Haberler*, Chicago 1965.

———, "More on the Equivalence of Tariffs and Quotas," *Amer. Econ. Rev.*, Mar. 1968, *58*, 142–146.

———, *Trade, Tariffs and Growth*, Cambridge, Mass. 1969.

H. Shibata, "A Note on the Equivalence of Tariffs and Quotas," *Amer. Econ. Rev.*, Mar. 1968, *58*, 137–42.

G. J. Yadav, "A Note on the Equivalence of Tariffs and Quotas," *Can. J. Econ.*, Feb. 1968, *1*, 105–10.

New York Times, Dec. 7, 1977.

Turnover and Mobility among the 100 Largest Firms: An Update

By ROBERT J. STONEBRAKER*

It is well known that aggregate concentration levels in the United States have been increasing over time (see Table 1). Although the economic importance of such concentration has long been controversial, this upward trend is, nonetheless, greeted by many with alarm (see, for example, John Blair, ch. 4). Since the potential dangers of aggregate concentration are likely to be magnified if the largest firms are solidly entrenched in their positions, an analysis of the turnover and mobility of these firms is of some import.

Several early analyses in this *Review* indicated that the degree of mobility among large firms was falling and that large firms were increasingly able to maintain their positions over time.[1] The purpose of this study is to reexamine the data and to determine if these trends have continued.

Section I reviews the results of the previous studies and concludes that the apparent diminution of turnover and mobility has ceased. Section II briefly considers some possible causes of the observed turnover and mobility, and Section III summarizes the results.

I. Trends in Turnover and Mobility

The most thorough study of the changing size structure of industrial giants was done by Collins and Preston (hereafter C-P). Building on some prior work by A. D. H. Kaplan, they identify the 100 largest *U.S.* firms in manufacturing, mining and distribution for the years 1909, 1919, 1929, 1935, 1948, and 1958. Firms are ranked in terms of asset size. The authors conclude that 1) there was no significant change in the distribution of assets *among* the 100 largest firms over the period, but 2) the amount of turnover as measured by the number of firms leaving the list of the top 100 or of new firms entering the list was declining, and 3) the amount of mobility of firms *within* the top 100 was decreasing over time. In a second study using somewhat different methodology, Mermelstein (1969) obtained similar results. Furthermore, after examining the asset shares of firms over time, he concluded that there was a definite increase in the the ability of the largest firms to maintain their relative positions. As Richard Caves has stated, such results are sometimes seen as evidence of "hardening of the industrial arteries and decreased competitiveness" (p. 40).

In order to update these studies, the list of the largest 100 firms in terms of assets has been calculated for the years 1967 and 1976. The asset data were derived from *Fortune* and all definitions are coincident with those used by C-P.

A. *Distribution of Assets among the Top 100 Firms*

To see if any change has occurred in the size distribution of assets among the largest firms, the percent of assets of the 100 largest firms held by the top 10, 25, 50, and 75 firms was calculated for each year (see Table 2). The stability noted by C-P has continued. Other than for a slight move toward equality from 1909 to 1919, there is no discernable trend in the relative importance of any group. If Lorenz curves were drawn for each of the eight observation years, they would be almost identical.

B. *Turnover among the Top 100 Firms*

For the entire 1909–76 period, the amount of turnover among the largest 100 firms has been significant. Only 21 of the top 100 firms

*Professor of economics, Indiana University of Pennsylvania. The research assistance of Gary S. Klonicki is gratefully acknowledged.

[1]See particularly N. R. Collins and L. E. Preston, and David Mermelstein (1969). Seymour Friedland reached similar conclusions as well.

TABLE 1—PERCENT OF TOTAL ASSETS CONTROLLED BY THE LARGEST FIRMS

Year	Largest 10 Firms	Largest 25 Firms	Largest 50 Firms	Largest 75 Firms	Largest 100 Firms
1909	7.4	10.7	14.0	16.1	17.7
1919	5.9	9.2	12.5	14.9	16.6
1929	9.1	14.6	19.6	23.0	25.5
1935	11.0	16.7	22.0	25.6	28.1
1948	9.8	15.6	20.7	24.2	26.7
1958	11.5	17.7	23.1	27.0	29.8
1967	11.5	18.4	24.6	28.7	32.0
1974	11.3	17.9	23.6	27.6	30.7

Sources: Asset data and rankings from 1909–58 are taken from Collins and Preston. Firm data for 1967 and 1974 are calculated from the "Directory of the 500 Largest U.S. Industrial Corporations," June 15, 1968, pp. 186–207; May 1975, pp. 208–35, and "Directory of the Fifty Largest Commercial-Banking Companies . . . ," June 15, 1968, pp. 208–21; July 1975, pp. 114–27. Total corporate assets for 1967 and 1974 are from *Statistics of Income; Corporate Income Tax Returns*. Data include all firms in mining, manufacturing, and distribution.

TABLE 2—PERCENT OF ASSETS OF THE TOP 100 FIRMS HELD BY THE TOP 10, 25, 50, AND 75 FIRMS

Year	Percent Held by Largest 10	Percent Held by Largest 25	Percent Held by Largest 50	Percent Held by Largest 75
1909	42.2	60.3	79.1	91.1
1919	35.5	55.2	75.4	89.7
1929	35.6	57.1	77.0	90.3
1935	39.1	59.2	78.1	91.2
1948	36.6	58.4	77.6	90.7
1958	38.7	59.4	77.5	90.6
1967	36.1	57.5	77.0	89.8
1976	36.4	58.4	77.8	90.3

Sources: Data from 1909–58 are taken from Collins and Preston. The 1967 and 1976 data are calculated from "Directory of the 500 Largest U.S. Industrial Corporations," June 15, 1968, pp. 186–207, May 1977, pp. 364–94, and "Directory of the Fifty Largest Commercial-Banking Companies . . . ," June 15, 1968, pp. 208–21; July 1977, pp. 160–73.

TABLE 3—NUMBER AND ASSET SHARES OF NEW ENTRANTS TO THE LIST OF THE LARGEST FIRMS[a,b]

Period	Entrants to Top 100	Asset Share	Entrants to Top 75	Asset Share	Entrants to Top 50	Asset Share	Entrants to Top 25	Asset Share	Entrants to Top 10	Asset Share
1909–19	40	31.3	31	32.4	20	30.6	13	36.3	7	40.5
1919–29	31	18.5	27	21.6	17	20.6	8	20.8	4	26.9
1929–35	16	5.5	10	5.5	9	8.2	2	3.5	2	12.6
1935–48	19	12.2	16	9.3	8	7.5	5	11.1	3	20.6
1948–58	16	8.2	15	9.3	12	11.6	3	5.6	1	5.8
1958–67	21	12.6	17	13.5	12	14.6	6	12.4	3	21.5
1967–76	18	9.5	16	11.4	11	11.1	4	9.3	1	7.5

Sources. Collins and Preston, and "Directory of the 500 Largest U.S. Industrial Corporations," June 15, 1968, pp. 186–207; May 1977, pp. 364–94, and "Directory of the Fifty Largest Commercial-Banking Companies . . . ," June 15, 1968, pp. 208–21; July 1977, pp. 160–73.

[a]Asset share is calculated as the percent of assets of the group controlled by the new entrants to that group. Thus, for the 1909–19 period, the new entrants to the top 10 controlled 40.5 percent of the assets of the top 10.

[b]Asset shares are shown in percent

TABLE 4—SPEARMAN CORRELATION COEFFICIENTS FOR SIZE RANKING

	1909–19	1919–29	1929–35	1935–48	1948–58	1958–67	1967–76
Spearman Coefficient	0.65	0.70	0.89	0.83	0.79	0.81	0.80

in 1909 have remained on the list and only one (Exxon) has remained in the top 10 over the period. In fact, only three of the top 10 firms in 1976 (Exxon, Texaco, GE) were among the 100 largest in 1909.

However, the question is whether there is any evident trend in the amount of turnover. The data for each time period are shown in Table 3. Both the number of new firms entering the list during each time period and the percent of assets of the top 100 firms accounted for by the new entrants are given. Since the number 100 is of no special significance, the data are reported for new entrants to the top 10, 25, 50, and 75 firms as well.

The data indicate that turnover probably did decline at least over the 1909–29 period.[2] But since 1929 no trend is apparent. The decline noted in the C-P, Mermelstein, and Friedland studies has not continued. In virtually every category the turnover observed since 1958 is at least as great, and often greater, than that during the 1929 to 1958 periods.[3]

C. *Mobility within the Largest 100 Firms*

Following C-P we can also examine the amount of mobility within the group of the largest firms. Even if there were very few firms entering or leaving the top 100, there could still be significant changes in the rank ordering of those remaining on the list. For those firms remaining in the top 100 over each time period, a Spearman coefficient can be calculated showing the correlation between the rank orderings of the firms at the beginning and end of the period. The data (Table 4) indicate that the trend in mobility parallels that of turnover. The Spearman coefficients rise until 1929 (denoting decreased mobility), but then level off for the remainder of the periods at about 0.8 (showing stability).[4]

Thus, the C-P conclusions that both turnover and mobility among large firms are declining over time is no longer valid. There has been no observable trend in either for nearly fifty years.

Similarly, Mermelstein's finding that the largest firms are increasingly able to maintain their power over time no longer seems to hold. For each period he identified the "survivor" firms (i.e., those remaining in the top 100). He then ran a regression of the form:

$$FSH_i = a + bISH_i \quad (1)$$

where FSH_i and ISH_i are, respectively, the percent shares of total survivor assets held by the ith firm in the final year of the period and the initial year of the period. If the largest firms (those with assets above the mean for the group) increase their shares relative to the smaller firms, the coefficient b will exceed one. If the shares of the largest firms decline, b will be less than one. The calculated b coefficients are given in Table 5. The coefficients for the 1909–19 and 1919–29 periods are significantly less than one indicating a tendency for the largest firms to lose asset shares over those periods. But since 1929 the coefficients have been remarkably stable and reveal no statistical trend.[5]

Stanley Boyle has argued that Mermel-

[2] A large part of the turnover during the 1909–19 period, however, was due to the dissolution of the Standard Oil and American Tobacco trusts.

[3] Mermelstein (1969) has correctly noted that the results might be biased by using time periods of different length. To check on this, the results were recalculated on a per year basis. The adjusted data, however, yield the same conclusions and are not reported here.

[4] The Spearman coefficient for 1929–35 is probably biased upwards because of the shortness of the time period.

[5] To insure that there is no trend in the data not evident to the eye, a regression of the estimated bs as a function of time was run. No significant relationship was found for the 1929–76 period.

TABLE 5—CALCULATED REGRESSION ESTIMATES FOR b

Period	b Estimate	R^2
1909–19	0.7329	0.92
1919–29	0.6540	0.74
1929–39	0.9648	0.92
1939–48	0.9205	0.93
1948–58	1.1103	0.94
1958–67	1.0300	0.90
1967–76	0.9425	0.92

stein's methodology is potentially flawed by its exclusion of nonsurviving firms. To avoid this problem an alternative methodology was used. Since 1929 every firm in the top 25 at the beginning of a period was among the top 100 at the end of the period. As a result, the asset share of the original top 25 firms in each period can be compared to their share at the end of the period. If large firms are becoming increasingly entrenched, the ability of the top 25 firms to maintain their asset shares during the periods should be increasing over time. But this is not the case. The ratios of the asset share of the top 25 firms at the beginning of each period to the asset share of the same 25 firms at the end of each period are shown in Table 6. No discernible trend exists. The top 25 firms at the beginning of the later periods seem to have been no more successful in maintaining their asset share than were their counterparts in the earlier periods. We again conclude that mobility within the largest 100 firms has been stable over the last fifty years.

TABLE 6—ASSET SHARE CHANGES OF THE INITIAL TOP 25 FIRMS OVER EACH TIME PERIOD

Period	AS_0[a]	AS_1[b]	AS_0/AS_1
1929–35	0.571	0.585	1.025
1935–48	0.592	0.560	0.945
1948–58	0.584	0.582	0.996
1958–67	0.594	0.552	0.930
1967–76	0.575	0.566	0.985

Sources. Collins and Preston, and "Directory of the 500 Largest U.S. Industrial Corporations," June 15, 1968, pp 186–207; May 1977, pp. 364–94, and "Directory of the Fifty Largest Commercial Banking Companies . . ," June 15, 1968, pp. 208–21; July 1977, pp. 160–73.

[a] AS_0 is calculated as the assets of the top 25 firms as a percent of the assets of the top 100 firms in the initial year of the period.

[b] AS_1 is calculated as the assets of the same 25 firms as a percent of the assets of the top 100 firms in the final year of the period.

II. Causes of Turnover and Mobility

Although no detailed analysis of the causes of turnover and mobility will be attempted here, two possibly important factors can be readily identified: shifts in interindustry demands and mergers.

Friedland hypothesized that changes in the size distribution and rankings of giant firms simply reflected differences in the rates of growth of demand of their respective industries. Examining the 50 largest manufacturing firms from 1906–50 he concluded that about two-thirds of the differences in *firm* growth rates could be attributed to differences in *industry* growth rates.

To see if this relationship still holds, tests similar to Friedland's were performed for the largest 100 manufacturing firms for the 1954-72 period. Since reliable asset data by industry are not available, firms were ranked in terms of sales revenue. The 1954–72 period was chosen because of the availability of *Census of Manufactures* data. The top 100 firms in 1954 were obtained from the Fortune 500 list of industrial corporations and then traced, as much as possible, to 1972.[6] The growth of sales revenue for each firm over the period was then regressed against the growth of sales revenue for the industry to which it belonged. The results are

$$(2) \quad FG_i = 1.1 + \underset{(.110)}{0.836} IG_i \qquad R^2 = 0.41$$

where FG_i is the growth of sales revenues for the ith firm and IG_i is the growth of sales revenue for the industry to which the ith firm belonged for the 1954–72 period.[7] The number in parentheses is the standard error, so the coefficient is statistically significant at

[6] Of the original top 100 firms only 85 could be traced until 1972. The others were no longer identifiable, typically because of mergers.

[7] All sales revenue data were taken from the *Census of Manufactures*, 1954; 1972.

TABLE 7—ACTUAL AND PREDICTED AGGREGATE CONCENTRATION PATTERNS FOR 1954 AND 1972

	Share of Top 10[a]	Share of Top 25	Share of Top 50	Share of Top 75	Share of Top 100
Actual 1954	13.5	20.9	27 9	32.5	36.0
Predicted 1972[b]	14.8	22.1	28 8	33.0	36.3
Actual 1972	15.8	23.9	32.3	38.5	43 5

Sources: *Census of Manufactures*, 1954; 1972, and "Directory of the 500 Largest U.S. Industrial Corporations," July 1955, pp. 1-12; May 1973, pp. 220-44

[a]Share is calculated as the percent of total sales revenue in manufacturing accounted for by the respective number of firms.

[b]The predicted 1972 percents are those that would have occurred if the largest firms in 1954 had grown at the same rates as the industries to which they belonged.

the 99 percent confidence level. Industry growth is apparently still a major determinant of the size structure of large firms, but the R^2 of 0.41 is much lower than the two-thirds Friedland found in his earlier study.

A final test was to calculate the aggregate concentration that would have occurred in 1972 if the largest firms in 1954 had simply grown at the same rates as the industries to which they belonged and compare it with the actual 1972 data. The results are shown in Table 7. If each firm had grown at the same rate as its industry, the 1972 concentration patterns would have been nearly identical to those of 1954. Yet actual 1972 concentration data show a significant increase over 1954 levels. Industry growth and demand patterns alone cannot explain the increase in the relative size of the largest 10, 25, 50, 75, and 100 firms over the period. Further research could presumably determine if this increase could be explained by other factors such as the conglomerate merger wave of the 1960's.

III. Summary

During the early 1900's there was a significant drop in the amount of turnover and mobility among the largest 100 *U.S.* firms in manufacturing, mining, and distribution. Although several earlier studies concluded that this was a continuing trend, I have found that it is not. Every statistical measure of turnover or mobility examined has been stable since the late 1920's. Even though the relative importance of the largest 100 firms has apparently increased, these firms seem no more solidly entrenched or immune from competition than did their counterparts of fifty years ago.

REFERENCES

John M. Blair, *Economic Concentration Structure, Behavior and Public Policy*, New York 1972.

S. E. Boyle, "Large Industrial Corporations and Asset Shares: Comment," *Amer. Econ Rev.*, Mar. 1971, *61*, 163–67.

Richard Caves, *American Industry: Structure, Conduct, Performance*, 4th ed., Englewood Cliffs 1977.

N. R. Collins and L. E. Preston, "The Size Structure of the Largest Industrial Firms, 1909–1958," *Amer. Econ. Rev.*, Dec. 1961, *51*, 986–1011.

S. Friedland, "Turnover and Growth of the Largest Industrial Firms, 1906–1950," *Rev. Econ. Statist.*, Feb. 1957, *39*, 79-83

A. D. H. Kaplan, *Big Enterprise in a Competitive System*, Washington 1954; rev. ed., Washington 1964.

D. R. Kamerschen, "Large Industrial Corporations and Asset Shares: Comment," *Amer. Econ. Rev.*, Mar. 1971, *61*, 160–62.

James V. Koch, *Industrial Organization and Prices*, Englewood Cliffs 1974.

D. Mermelstein, "Large Industrial Corporations and Asset Shares," *Amer. Econ. Rev.*, Sept. 1969, *59*, 531–41.

———, "Large Industrial Corporations and Asset Shares: Reply," *Amer. Econ. Rev.*, Mar. 1971, *61*, 168–74.

F. M. Scherer, *Industrial Market Structure*

and Economic Performance, Chicago 1970.

Fortune, "The Directory of the 500 Largest U.S. Industrial Corporations," Chicago various issues.

———, "The Directory of the Fifty Largest Commercial-Banking Companies, Life Insurance Companies, Diversified-Financial Companies, Merchandising Companies, Transportation Companies, and Utilities," Chicago various issues.

Internal Revenue Service, *Statistics of Income*; *Corporate & Income Tax Returns*, Washington, various issues.

U.S. Bureau of the Census, *Census of Manufactures*, Washington 1954; 1972.

A Simple Neutrality Result for Movements between Income and Consumption Taxes

By John Whalley*

In this note the possibility is demonstrated that a movement between a broadly based income tax and a consumption tax in a two-period consumption loan model can be completely accommodated by interest rate changes which leave real intertemporal consumption plans unchanged. Income and consumption taxes are both broadly based taxes, the former taxing all potential consumption in any period and the latter actual consumption. Lenders and borrowers face the same prices under both tax regimes and movements between the two can, in this simple model, be wholly accommodated by interest rate changes leaving intertemporal consumption plans unaffected.

This result contrasts with the conventional argument in favor of a consumption tax in preference to an income tax on the basis of lack of distortion of savings behavior. It is not suggested that because of this result exact monetary accommodation to consumption income tax variations will occur in all circumstances, but it seems to be of interest to note that such adjustments are possible and these appear not to have been previously considered.

The traditional argument for the distorting effects of an income tax over a consumption tax is often made in a simple two-period intertemporal consumption choice model. If an individual receives income Y_1, Y_2 in each of two periods and if the interest rate is r, then, so the argument goes, the slope of an individual's budget constraint between current and future consumption (C_1 and C_2) is not disturbed by a consumption tax, whereas it is under an income tax. If interest is both taxable as a receipt and deductible as an expense under the income tax, and the marginal tax rate t is assumed to apply under both the income and consumption tax,[1] the slopes of the consumer budget constraint under the three alternative regimes are

(1) No Tax

$$\frac{P_1}{P_2} = (1 + r)$$

Consumption Tax (CT)

$$\frac{P_1}{P_2} = \frac{(1 + r)(1 - t)}{(1 - t)}$$

Income Tax (IT)

$$\frac{P_1}{P_2} = \frac{(1 + r(1 - t))(1 - t)}{(1 - t)}$$

where P_1 and P_2 refer to the prices of a unit of a homogeneous consumption good in the two periods. The slope of the budget constraint in the consumption tax case is equal to that of the budget constraint in the no-tax case and for this reason the consumption tax is regarded as a nondistorting tax.[2] With the income tax the substitution effect works in favor of increased current consumption and reduced future consumption (savings) and the tax is therefore regarded as a distorting tax.

*Department of economics, University of Western Ontario. I am grateful to the managing editor and an anonymous referee for helpful comments

[1]A characteristic of the consumption tax noted in a number of places (see, for instance, A R. Prest) is that with positive net savings in the economy equality of tax yields under the two taxes on a year-by-year basis will require a higher tax rate under the consumption tax in the initial year in which the capital stock of the economy is unchanged

[2]This argument only holds in the absence of a labor-leisure choice. With either a consumption or an income tax, leisure is an untaxed good, and the choice between these taxes reduces to counting and evaluating distortions (labor-leisure under the consumption tax, labor-leisure and savings under the income tax). Conditions will exist where an income tax dominates a consumption tax, although Anthony Atkinson and Joseph Stiglitz show that with homothetic preferences which are separable into leisure and nonleisure goods a consumption tax is an optimal tax given the inability to tax leisure.

This conventional argument makes no allowance for endogenous determination of r, and the main point of this note is to argue that conditions exist under which it is possible for r to change in such a way as to leave the slopes of budget constraints unaffected as movements between tax regimes occur. If interest rate changes from movements between tax bases are taken into account, equations (1) no longer serve as a basis for comparison between tax regimes. Additional equations are needed for the determination of r.

If r^{CT} and r^{IT} refer to the interest rates determined under the two tax regimes, the interesting case is where $(1 + r^{CT}) = (1 + r^{IT}(1 - t))$. In this event, movements between the tax bases will not change the relative prices of consumption in the two time periods, and if there are no additional income effects, there is no effect on savings which will occur through the introduction of a consumption tax

This result can be shown to hold in a simple two-period consumption loan model The argument is similar to that underlying the long-run neutrality of movements between origin and destination bases for general taxes. In this model, the adoption of a consumption tax in long-run equilibrium will have no effect on savings behavior in much the same way that movements between origin and destination bases for broadly based taxes leave the real characteristics of international trade unaffected.

Consider two individuals A and B, each of whom has an income stream covering two periods (Y_1^A, Y_2^A), (Y_1^B, Y_2^B). Each has a utility function defined on consumption in each period:

$$U^A(C_1^A, C_2^A) \qquad U^B(C_1^B, C_2^B) \tag{2}$$

and each determines a desired consumption stream on the basis of utility maximization subject to the intertemporal budget constraint obtained by discounting second period income at the interest rate they face.[3]

Because of differences in preferences and income profiles each individual will typically desire a consumption stream different from his income stream. In equilibrium, therefore, one individual will "lend" first-period consumption to the other individual, receive an interest return on his savings, and consume principal and interest (repaid by the other individual) in the second period along with period-two income. The analogy to a pure international trade model therefore is clear, trade occurs between periods rather than between countries.[4]

Consider the no-tax situation and suppose, for simplicity, that in equilibrium individual A is the saver in period one. The equilibrium conditions which determine r are that in period one, savings by A equal borrowings by B and in period two, dissavings by A equal repayments by B; and that for both individuals their desired consumption streams are solutions from their utility-maximization exercises for the equilibrium value of r. These conditions may be written as

$$(Y_1^A - C_1^A) = (Y_2^B - C_2^B)/(1 + r) \tag{3}$$

$$(C_1^A, C_2^A) = f^A(1 + r) \tag{4}$$
$$(C_1^B, C_2^B) - f^B(1 + r)$$

If the introduction of a consumption tax at a rate t^C is now considered with the assumption on tax revenues that they are returned in lump sum manner so as to leave the real income stream of each individual unchanged these conditions become

$$(Y_1^A + t^C C_1^A - (1 + t^C)C_1^A = (Y_2^B + t^C C_2^B - (1 + t^C)C_2^B)/(1 + r) \tag{5}$$

$$(C_1^A, C_2^A) = f^A(1 + r) \tag{6}$$
$$(C_1^B, C_2^B) = f^B(1 + r)$$

It follows that as (5) and (6) reduce to (3) and (4) equilibrium under the consumption tax can be achieved with an unchanged in-

[3]It is assumed that individuals can both borrow and lend at the same interest rate.

[4]The result obtained below does not appear to hold in an ongoing consumption loan model with overlapping generations, as considered by Paul Samuelson. In the Samuelson model, workers save for retirement on the basis of the anticipated interest rate to be paid by future generations. Trade takes place between current and future working populations and is unidirectional rather than reciprocal

terest rate r and unchanged consumption streams (C_1^A, C_2^A), (C_1^B, C_2^B). The tax is nondistorting; no real characteristics of intertemporal equilibrium are affected. The same equilibrium is achieved independently of the tax rate since no distorting properties of the tax are present.

The introduction of an income tax at a rate t^I can also be considered under the same revenue assumption that they are returned in lump sum manner so as to leave real income streams unchanged. Because of the income tax paid on interest received in period two, period-two tax collections under the income tax are higher if $t^I = t^C$. As all tax proceeds are returned in lump sum manner so as to preserve the gross of tax income streams, any income effects which would be associated with proportional division of tax revenues in each period are removed. The equilibrium conditions in the presence of the income tax, analogous to (3) and (4), are

$$(7)\quad (Y_1^A - C_1^A) = (Y_2^B - C_2^B)/(1 + r(1 - t^I))$$

$$(8)\quad (C_1^A, C_2^A) = f^A(1 + r(1 - t^I))$$
$$(C_1^B, C_2^B) = f^B(1 + r(1 - t^I))$$

The term $(1 + r(1 - t^I))$ appears as the discounting factor in (7) since B receives a tax deduction in period one on interest paid. While A must dissave to pay the tax on interest received, this tax paid appears as part of the lump sum redistributed tax revenue and thus does not necessitate foregone consumption by B in period two.

Let r^{CT} and r^{IT} denote the equilibrium interest rates achieved under the consumption and income taxes, respectively. It was seen that the equilibrium r^{CT} equals the equilibrium interest rate with no tax. It is clear from comparing (7) and (8) to (3) and (4) that r^{IT} must adjust to restore equilibrium. In equilibrium, $r^{IT}(1 - t^I)$ must equal the no-tax interest rate. Savings behavior will remain unaffected by a movement from a consumption tax to an income tax or vice versa. This result applies irrespective of the tax rates involved, since the income effects associated with different levels of tax collections are removed from the analysis. The analogy to origin vs. destination basis results is evident With trade across periods rather than countries the interest rate rather than the exchange rate can adjust to offset any anticipated real effects of the tax change.

REFERENCES

A. B. Atkinson and J. E. Stiglitz, "The Design of Tax Structure: Direct Versus Indirect Taxation," *J. Publ. Econ.*, July/Aug. 1976, *16*, 55–75.

A. R. Prest, *Public Finance in Theory and Practice*, London 1975.

P. A. Samuelson, "An Exact Pure Consumption Loan Model of Interest With or Without the Contrivance of Money," *J. Polit. Econ.*, Dec. 1958, *66*, 467–82.

Substitution vs. Addiction in the True Index of Real Wages

By LOUIS PHLIPS AND PIERRE PIERAERTS*

In a recent article in this *Review*, John Pencavel (1977) convincingly criticized the current practice of evaluating changes in real wages by comparing changes in money wages with changes in an index of consumer goods' prices, on the grounds that this practice underestimates the rise in real wages. The bias results from the fact that the official price indexes use base-year weights, that is, of the Laspeyres type, and therefore overstate increases in the cost of living. By reintroducing the celebrated substitution effect—not only between commodities but also between commodities and leisure—one can show that base-year utility can be maintained with a lower wage rate than the one measured by the Laspeyres approach. Pencavel does even more: as the problem is set up in the framework of a model of the allocation of time, he is also able to treat the wage rate as a source of income (and not only as the price of the commodity "leisure") distinct from nonlabor income.

A natural question to ask is to what extent the utility-maximization approach on which constant utility indexes are based is itself biased as a result of the qualification "tastes remaining the same." What if the average worker to whom Pencavel's computations refer is addicted to certain goods, so that his/her marginal utilities rise over time as a result of increased consumption? What if, in particular, leisure is an "addictive" good, with the result that the exposure to longer holidays increases the subsequent demand for holidays? If the phenomenon can be tackled theoretically and empirically, one would not be surprised to find that the rise of real wages, as measured on the basis of given (base-year) tastes, in fact overestimates the gain as perceived by the worker who is subject to addiction.

In an article in the same issue of this *Review*, George Stigler and Gary Becker accumulate examples to show that taste changes such as those alluded to can be handled theoretically with the usual utility-maximizing apparatus. To do so, it suffices to make the arguments of the utility function dependent upon past consumption in order to represent the accumulation of what is termed "consumption capital." As a result, marginal utilities change over time while the utility function itself remains unchanged. We want to recall that the state variable approach, pioneered by Hendrik Houthakker and Lester Taylor and further developed by Phlips (1974, 1977, 1978), proceeds exactly along those lines on *both* a theoretical and empirical level.

This paper follows the state variable approach to define two families of taste-dependent[1] constant utility indexes of real wages (Section I). We then "dynamize" the Stone-Geary utility function and derive the implied taste-dependent true wage rates (Section II). Using the same data set as Pencavel (1979) we compare our dynamic indexes with his static index (Section III), to conclude that a neglect or an incorrect handling of taste changes indeed biases constant utility indexes of real wages upwards. The currently used official index numbers of real earnings also turn out to be biased upwards, because the

*Professor of economics, CORE, and research assistant, Institut des Sciences Economiques, Université Catholique de Louvain, respectively. We are grateful to J-P Lemaître and D. Van Grunderbeeck for computational assistance. O. Ashenfelter and J. Pencavel provided much appreciated comments, data and insights, while A. P. Barten kindly discussed the manuscript. All remaining errors are ours.

[1]When cash balances held for transactions purposes are included in the analysis, it is possible to enlarge the approach even further by embedding the "full-income" constraint in a wealth constraint defining savings as shown in Phlips (1977, 1978). Although savings do not appear here explicitly, they are not supposed to be zero, in contrast with the static theory of the allocation of time.

addiction effect appears to be stronger than the substitution effect.

I. Taste-Dependent True Wage Indexes

Suppose the average worker maximizes the dynamized utility function

$$u = u(x, l; s, s_l) \tag{1}$$

where x is an n-vector of commodity purchases, l represents hours of leisure, and $u_i > 0$, $u_l > 0$. The variable s is an n-vector of state variables, each associated with a corresponding quantity purchased, while s_l is the state variable associated with leisure. These states designate what Stigler and Becker call consumption capital, as they are defined as the solutions of

$$ds_i/dt = x_i - \delta_i s_i \tag{2a}$$

$$ds_l/dt = l - \delta_l s_l \tag{2b}$$

respectively, δ being the rate of depreciation (i.e., the rate at which addiction wears off). In the Houthakker-Taylor terminology, the states are called "stocks of habits" and addiction is called "habit formation." The utility function (1) is maximized subject to the constraint

$$y = \sum p_i x_i + p_l l \tag{3}$$

which defines full income. Insertion of the demand equations into the utility function gives the indirect utility function

$$u = u(p, p_l, y; s, s_l) \tag{4}$$

where p is an n-vector of prices and p_l is the wage rate.

Taste changes are captured through the presence of the states—as parameters—in the utility function: to the extent that the states change over time, marginal utilities shift while (1) or (4) remain valid representations of the preference ordering at each point in time. Any purchase (of a unit of a commodity or of an hour of leisure) today changes the values of the states tomorrow through (2a) and (2b), which in turn implies a new utility level, and so on. Yesterday's prices and income determine yesterday's purchases, which determine today's utility function and therefore today's purchases.... Furthermore, yesterday's consumption determines today's labor income, given that hours of leisure and working hours (h) are linked by the equality $l = T - h$, where T is the (unknown) maximum numbers of hours to be allocated.

To define a taste-dependent true wage, we shall proceed in two ways. We shall first define what we shall call the cardinal true wage rate p_{lt}^*. (It is "cardinal" because a comparison between tastes in the base year and tastes in the current year t is involved.)[2] We shall then define the Fisher-Shell (hereafter F-S) true wage rate p_{lt}^{**}, which is "ordinal" to the extent that reference is made to current tastes only to avoid any intertemporal utility comparison.

The cardinal constant utility wage rate p_{lt}^* is defined as the solution of

$$u(p_0, p_{l0}, y_0; S_0, s_{l0}) = u(p_t, p_{lt}^*, y_t; s_t, s_{lt}) \tag{5}$$

where the subscript 0 designates the base year. In accordance with Pencavel's terminology, the cardinal true wage index is then p_{lt}^*/p_{l0}, while the cardinal real wage index is p_{lt}/p_{lt}^*. When the latter is larger than one, the average worker is better off than in the base year given the change in tastes between 0 and t.

The F-S constant utility wage rate p_{lt}^{**} may be defined as the solution of

$$u(p_0, p_{l0}, y_0; s_t, s_{lt}) = u(p_t, p_{lt}^{**}, y_t; s_t, s_{lt}) \tag{6}$$

in which only *current* tastes (represented by s_t and s_{lt}) are taken into account to follow "ordinal" orthodoxy as advocated by F-S. The corresponding indexes may then be called F-S true and real wage indexes. When the F-S real wage index p_{lt}/p_{lt}^{**} is larger than one, the average worker is better off than he would have been if today (given today's

[2]The adjective cardinal is, of course, a misnomer to the extent that a) p_{lt}^* is invariant under monotonic increasing transformations of the utility function, and b) the comparison of tastes over time can be reinterpreted as a comparison of alternative (past) time paths of the state variables, i.e., of alternative amounts of consumption capital

tastes) he had been confronted with base-year prices and base-year full income. Just as the F-S true income (that defines the numerator of the true price index) is below the cardinal true income in case of addiction, the F-S true wage is lower than the cardinal true wage when commodities are habit forming.

II. Specification of the Utility Function

In order to obtain results comparable with Pencavel's, we shall also use the Stone-Geary utility function. However, to introduce taste changes, we shall generalize it by making the minimum subsistence quantities depend (linearly) upon the corresponding state variables, i.e.,

$$(7)\quad u = \sum_i \beta_i \ln (x_i - \gamma_i) + \beta_l \ln (l - \gamma_l) \qquad (i = 1, \ldots, n)$$

with

$$(8)\quad \gamma_i = \theta_i + \alpha_i s_i$$
$$\gamma_l = \theta_l + \alpha_l s_l$$

When α is positive the taste change is "quantity diminishing", there is habit formation or addiction. (It is understood that the state variables are positive quantities.)

Maximization of (7) subject to (8) and the constraints (2) and (3) leads to the system of estimating equations.[3]

$$(9)\quad p_{it}x_{it} = p_{it}\Big(K_{i0} + K_{i1}x_{it-1} + K_{i2}\frac{1}{\lambda_t p_{it}} + K_{i3}\frac{1}{\lambda_{t-1}p_{it-1}}\Big)$$
$$p_{lt}(-h_t) = p_{lt}\Big[K_{h0} + K_{h1}(-h_{t-1}) + K_{h2}\frac{1}{\lambda_t p_{lt}} + K_{h3}\frac{1}{\lambda_{t-1}p_{lt-1}}\Big]$$

The cardinal constant utility wage rate p_{lt}^* is simply the solution of

$$(10)\quad 1 = \frac{z_0 + p_{l0}\gamma_{h0} - \sum_i p_{i0}\gamma_{i0}}{z_t + p_{lt}^*\gamma_{ht} - \sum_i p_{it}\gamma_{it}} \prod_i \Big(\frac{p_{it}}{p_{i0}}\Big)^{\beta_i}\Big(\frac{p_{lt}^*}{p_{l0}}\Big)^{\beta_l}$$

and can easily be computed once estimates of γ_{it} and γ_{ht} are available. Thus $\gamma_h = T - \gamma_l$ is the *maximum* number of hours one is willing to work and should decrease over time if leisure turns out to be addictive ($\alpha_l > 0$). The variable z is equal to $\Sigma_i p_i x_i + p_l(-h)$.

The F-S constant utility wage rate p_{lt}^{**} obeys the same formula (10) once γ_{i0} and γ_{h0} are replaced by γ_{it} and γ_{ht}, respectively.

III. Dynamic vs. Static Indexes

Using Michael Abbott and Orley Ashenfelter's corrected data (see Pencavel, 1979), we obtain[4] the regression results reported in Table 1. All regression coefficients have the expected sign. Except for three K_0 coefficients, they are highly significant. And to the extent that residuals are not autocorrelated (which seems to be the case according to the first-order autocorrelation coefficients ρ_i reported in the last column of Table 1), they are consistent and asymptotically efficient. The demand for "other services" is unstable ($K_1 > 1$) as in earlier work.

These regression coefficients are used to compute the structural coefficients listed in Table 2. Most commodity groups including leisure seem to be addictive, but transportation does not display any dynamic behavior. The other structural parameters are reported for comparison purposes.

As the emphasis is on the demand for leisure, it is of interest to take a closer look at its parameters. First of all, the time shape of γ_{ht} (see Table 3) displays interesting features. The overall trend downwards reflects the overall increase in the minimum required

[3]The derivation of this system and the estimation procedures are fully discussed in Phlips (1978). The computation of γ_{it} and γ_{ht} is also explained there. The marginal utility of z is λ_t. The K_i and K_h regression coefficients are constants and functions only of the β, θ, α, and δ structural coefficients.

[4]The results were obtained after seventeen iterations on λ_t (the marginal utility of z) made to impose the budget constraint. At that point, the relative difference between observed and estimated z_t was at most 0.002 (except for the years 1942, 1943, and 1944 for which it was 0.007).

TABLE 1—REGRESSION COEFFICIENTS

Group	K_{i0}	K_{i1}	K_{i2}	K_{i3}	ρ_i
Durables	−0.03	0.696	200.90	−95.38	−0.04
	(0.04)	(0.035)	(22.24)	(23.59)	(0.16)
Food	0.24	0.793	167.22	−130.68	−0.06
	(0.03)	(0.023)	(9.59)	(9.82)	(0.17)
Clothing	−0.02	0.981	56.22	−47.52	−0.23
	(0.04)	(0.057)	(7.46)	(8.44)	(0.17)
Other Nondurables	0.07	0.883	71.12	−42.18	0.08
	(0.01)	(0.019)	(4.61)	(4.90)	(0.17)
Housing Services	0.03	0.918	73.21	−32.19	−0.05
	(0.01)	(0.011)	(8.76)	(9.33)	(0.16)
Transportation Services	0.06	0.952	19.51	−18.79	−0.09
	(0.05)	(0.071)	(3.58)	(3.48)	(0.17)
Other Services	−0.04	1.022	74.41	−64.16	−0.12
	(0.02)	(0.018)	(9.53)	(9.83)	(0.17)
Supply of Labor	−392.40	0.847	90.71	−42.73	−0.01
	(62.81)	(0.027)	(21.13)	(19.34)	(0.17)

TABLE 2—STRUCTURAL COEFFICIENTS

Group	β_i	$\beta_i/\Sigma\beta$	θ_i	α_i	δ_i
Durables	174.1	0.261	−0.058	0.353	0.712
	(25.9)		(0.065)	(0.124)	(0.144)
Food	166.2	0.248	1.085	0.014	0.245
	(10.4)		(0.154)	(0.027)	(0.043)
Clothing	52.4	0.078	−0.096	0.148	0.168
	(7.4)		(0.291)	(0.051)	(0.087)
Other Nondurables	60.2	0.089	0.142	0.387	0.511
	(4.6)		(0.026)	(0.066)	(0.083)
Housing Services	55.0	0.082	0.037	0.692	0.778
	(9.2)		(0.018)	(0.155)	(0.161)
Transportation Services	19.6	0.029	1.633	−0.012	0.038
	(3.4)		(5.766)	(0.089)	(0.152)
Other Services	68.5	0.102	−0.242	0.170	0.148
	(9.2)		(0.153)	(0.065)	(0.073)
Allocation of Hours	72.3(β_l)	0.108	591.0(θ_h)	0.553(α_l)	0.719(δ_l)
	(20.9)		(202.5)	(0.250)	(0.253)

hours of leisure γ_l. The business cycle is also present. In particular, γ_h reaches a peak in 1944, when the war effort was at its highest, and then starts to decline, following the evolution of the number of hours worked. During all years, $\gamma_{ht} - h_t$ (as well as $x_{it} - \gamma_{it}$, all i) is positive with the implication that the utility interpretation of our results is valid for the entire observation period.

The supply of labor turns up with a (small) negative uncompensated price elasticity in Table 4: the supply of labor is slightly backward bending. Table 4 also reveals that a change in the wage rate, through its impact as a source of income, strongly affects the demand for durables. A compensated wage change, however, has almost no impact.

Table 5 gives observed rates after taxes (as computed by Abbott and Ashenfelter) for selected years, the static estimates of the true wage rate (as reported by Pencavel, 1979), and our two dynamic estimates. Table 6 lists the corresponding true and real indexes.

Looking at the cardinal constant utility wages first, we see that they rise less than observed wages and that the difference is

TABLE 3—HOURS OF WORK

Year	h_t	γ_h	$\gamma_h - h_t$
1929	2579	2766	87
1930	2530	2668	138
1931	2494	2614	120
1932	2409	2521	112
1933	2395	2488	93
1934	2210	2295	85
1935	2260	2343	83
1936	2326	2407	81
1937	2372	2448	76
1938	2297	2387	90
1939	2334	2422	88
1940	2340	2424	84
1941	2361	2427	66
1942	2416	2444	28
1943	2465	2479	14
1944	2489	2513	24
1945	2427	2488	61
1946	2308	2437	129
1947	2252	2370	118
1948	2228	2335	107
1949	2223	2335	112
1950	2197	2309	112
1951	2185	2274	89
1952	2187	2274	87
1953	2159	2251	92
1954	2139	2241	102
1955	2161	2266	105
1956	2151	2251	100
1957	2121	2223	102
1958	2099	2207	108
1959	2122	2228	106
1960	2126	2230	104
1961	2110	2214	104
1962	2117	2221	104
1963	2117	2223	106
1964	2122	2228	106
1965	2134	2240	106
1966	2126	2227	101
1967	2126	2223	97

TABLE 5—WAGE RATES (*U.S.* DOLLARS PER HOUR)

	Observed p_h	Static w_t^*	Cardinal p_h^*	F-S p_t^{**}
1946	0.709	0.709	0.709	0.709
1950	0 972	0.796	0.974	0.795
1955	1.288	0.932	1.231	0.917
1960	1.607	0 934	1.489	0.936
1965	1.985	0.888	1.709	0.912
1967	2.181	0 908	1.929	1.012

increasing. As a result the cardinal real wage index p_h/p_h^* indicates an improvement of about 13 percent in 1967 as compared with 1946. This is much below the 125 percent increase in the static index, and also below what the Laspeyres-type indexes of real earnings (such as the *BLS* indexes) indicate. Obviously, habit formation has more than neutralized the gains due to the substitution effect.

The F-S index climbs much higher than the cardinal index and stays rather close to the static index (see Table 6). This is not surprising as the F-S index refers only to the (ever changing) *current* preference ordering; by so doing it ignores taste changes to the extent that they cause displacements of indifference curves. As a result (as shown elsewhere)[5] the F-S index penalizes the worker who consumes addictive goods (their indifference curves may be said to "shift upwards"). In the present context, this implies that the true

[5]See Phlips (1974, ch. 9) for a graphical proof, and Eric De Souza for a formal demonstration.

TABLE 4—SHORT-RUN ELASTICITIES (SIMPLE AVERAGES)

Group	Compensated Price	Uncompensated Price	Compensated Wage	Uncompensated Wage
Durables	−0.42	−0.68	0.06	1.67
Food	−0.18	−0.43	0 02	0.64
Clothing	−0 19	−0.27	0.02	0.54
Other Nondurables	−0.16	−0.26	0.01	0.49
Housing Services	−0.10	−0 18	0.01	0.31
Transportation Services	−0.23	−0.26	0.02	0 65
Other Services	−0 15	−0.26	0.01	0.48
Supply of Labor	0.03	−0.07	0.03	−0.07

TABLE 6—WAGE INDEX NUMBERS

	Static w_t^*/w_0	Cardinal p_{lt}^*/p_{l0}	F-S p_{lt}^{**}/p_{l0}	Static w_t/w_t^*	Cardinal p_{lt}/p_{lt}^*	F-S p_{lt}/p_{lt}^{**}
1946	1.000	1 000	1 000	1.000	1.000	1.000
1950	1.123	1 373	1 121	1 221	0.998	1.222
1955	1.315	1 736	1 294	1 382	1.046	1 404
1960	1.317	2.100	1 320	1.721	1.080	1.717
1965	1.252	2.410	1.287	2.235	1.161	2.175
1967	1.365	2 721	1.428	2 253	1.131	2.154

wage p_{lt}^{**} does not increase as much as it does in the cardinal index, in which displacements of indifference curves are duly reflected.

To sum up: a) a static utility-maximization approach to the measurement of real earnings exaggerates the gain in the real wages; b) the same is true in a dynamic approach when the F-S index, focusing on current tastes, is used and goods consumed are addictive; c) to the extent that addiction is properly taken into account, as is the case in our cardinal index, the gain in real wages is even smaller than the one indicated by the currently used Laspeyres-type *BLS* index numbers.[6]

[6]Another commodity breakdown, used in Phlips (1977), also indicates that the gain in real wages is smaller than what the *BLS* indexes suggest

REFERENCES

M. Abbott and O. Ashenfelter, "Labor Supply, Commodity Demand, and the Allocation of Time," *Rev. Econ. Stud.*, Oct. 1976, *43*, 389–411.

E. De Souza, "Taste Change in the True Cost-of-Living Index," *Rech. Econ. Louvain*, No. 1, 1974, *40*, 55–68.

F. M. Fisher and K. Shell, "Taste and Quality Change in the Pure Theory of the True Cost-of-Living Index," in J. N. Wolfe, ed., *Value, Capital and Growth, Essays in Honour of J. R. Hicks*, Oxford 1969.

Hendrik S. Houthakker and Lester D. Taylor, *Consumer Demand in the United States 1929-1970*, 2d ed., Cambridge, Mass. 1970.

J. H. Pencavel, "Constant-Utility Index Numbers of Real Wages," *Amer. Econ. Rev.*, Mar. 1977, *67*, 91-100.

———, "Constant-Utility Index Numbers of Real Wages: Revised Estimates," *Amer. Econ. Rev.*, Mar. 1979, *69*, 240–43

Louis Phlips, *Applied Consumption Analysis*, Amsterdam 1974.

———, "A Taste-Dependent True Wage Index," in Wolfgang Eichhorn et al., eds., *Theory and Applications of Economic Indices*, Würzburg-Wein 1977.

———, "The Demand for Leisure and Money," *Econometrica*, Sept. 1978, *46*, 1025–43.

G. J. Stigler and G. S. Becker, "De Gustibus Non Est Disputandum," *Amer. Econ. Rev.*, Mar. 1977, *67*, 76–90.

Minimizing AEA Convention Transportation Costs

By JOHN J. SIEGFRIED AND LARRY NELSON*

There are probably more self-appointed experts on the optimal location of the annual American Economic Association (AEA) convention than on any other Association issue. Opinion is divergent because of the differing interests and tastes of the membership as well as the absence of agreement on the objective function that is relevant to determining the convention location. The diversity of employers' travel reimbursement policies may also contribute to disagreement on the optimal location of the AEA convention.

The annual convention provides an opportunity to present research findings to one's professional colleagues and receive (sometimes) constructive criticism, observe the frontier of economic research by listening to papers that will be published only after considerable delay, meet old friends (and enemies), discuss professional issues, generate research ideas, relive the "good old days," gossip about the profession, sometimes have one's institution pay for a vacation that would have been taken anyway, participate in the labor market for economists, examine new books first hand at the publishers' exhibits, honor distinguished colleagues, and occasionally have egos inflated or crushed.[1] The AEA convention location is a compromise that presumably maximizes some unspecified net benefits function. Gross benefits vary by location because of differences in tastes for various amenities and environmental attributes (for example, swimmers may prefer Miami to Chicago while theater-goers would most likely prefer New York City to Dallas) and because of differences in budget constraints among AEA members. Gross costs include transportation expenditures, lodging and meals, and the opportunity cost of time in attendance at the convention.

The greatest variation in costs undoubtedly arises from differences in transportation expenses. Nevertheless, it is sensible to choose a convention location with higher than minimum transportation costs if the incremental transportation costs are outweighed by the sum of other cost savings (for example, hotel charges) and additional gross benefits (for example, entertainment value). In order to aid in assessing the various incremental costs and benefits of alternative convention locations, we have computed approximate relative transportation costs of the AEA membership to twenty-six major North American cities.[2] We recognize that certain other costs may be important in the decision of an individual to

*Associate professor of economics, Vanderbilt University, and assistant professor of economics and finance, Middle Tennessee State University, respectively. John L Bungum, Rendigs Fels, C Elton Hinshaw, George Sweeney, and an anonymous referee provided helpful comments on an earlier draft, and O. Lee Pearson assisted in computer programming The Vanderbilt University Research Council provided financial support for some data collection.

[1]A recent survey of conferees at four professional business conventions reported an average of 46 percent of waking hours spent in professional activities such as meetings, seminars, and discussions with professional colleagues, 30 percent of the time spent in social activities, and 21 percent of hours spent in seeking employment or recruiting. These values obviously have a considerable variance among individuals, especially the employment and recruiting figures. The same survey reported that the extent to which expenses are reimbursed and the geographic location of the conference were the third and fifth most important elements in the decision to attend. Attendance by professional colleagues, the program topics or format, and the individuals' participation on the program were the first, second, and fourth most frequently reported reasons for attending. For more details see Leete A. Thompson, Ralph M Gaedeke, and Dennis H. Tootelian, pp. 1-3.

[2]The cities were suggested to us by the Secretary of the AEA, who is responsible for identifying and evaluating alternative convention sites. These twenty-six cities were suggested based upon estimates of available hotel space and of sufficient airline service to accomodate the AEA convention. A few cities which might have met these criteria have not been included because of the extremely low probability of their hosting the annual AEA convention (for example, Mexico City).

attend the convention; however, transportation expenses vary more with location than do the other important components of total cost.

I. Cost-Minimizing Algorithm

If we assume that the geographic distribution of AEA conventioneers is similar to the geographic distribution of the entire membership,[3] that all economists traveling from a given location to the convention incur equivalent *unit* transportation costs, and that travel is along a straight line from members' home airports[4] to the convention location, the preferred method of ascertaining the location at which aggregate member travel costs are minimized can be expressed as

$$\text{Min } TTC = \sum_{i=1}^{n} E_i C_i [(\bar{x} - x_i)^2 + (\bar{y} - y_i)^2]^{1/2}$$

where TTC = total transportation costs
E_i = the number of economists at location i
C_i = the transportation cost per economist per mile
$x_i; y_i$ = the latitude and longitude of location i
$\bar{x}; \bar{y}$ = the latitude and longitude of the convention location
i = economists' home location airports
n = number of home location airports.

This problem is analogous to a standard question in production management: Where is the optimal location of a single distribution center with known fixed-demand quantities and locations?[5] When C_i is assumed constant across all i, TTC produces the minimum total travel mileage for conventioneers.

The ease with which the problem can be stated belies the difficulty in solving it. The problem would be simplified considerably if travel were restricted to east-west and north-south movements, in which case the cost-minimization location would correspond to the intersection of (a) the latitude coordinate of the median economist if all were distributed along a single north-south line, and (b) the longitude coordinate of the median economist if all were distributed along a single east-west line. Travel along a direct line complicates the problem, but is a necessary assumption in view of the common use of air travel to go to conventions.

There are several feasible but lengthy algorithms that have been devised to solve the transportation cost-minimization problem analytically (see Zimmermann and Sovereign). One particular method also provides an intuitive understanding of the solution algorithms. First, drill frictionless holes in a solid fixed surface, each hole corresponding to a pair of coordinates x_i and y_i. Then suspend frictionless strings of equal length through all the holes and fasten weights representing the number of economists at each location to the respective string for that location. Tie all of the strings together in a large knot on top of the surface, and drop the knot. If the knot does not go through a hole, its location is the minimum transportation cost location of the AEA convention (assuming a geographical distribution of attendance proportional to the geographical distribution of membership, constant unit travel costs, and transportation on a direct line between home and convention

[3]Although this assumption is not likely true with respect to an actual convention because of higher attendance rates of people living closer to the convention, it still makes sense to consider all members' preferences and locations when planning a convention site. We assume that the greater propensity of nearby members to attend occurs equally for each potential location, so the bias in our estimate will be slight (but not zero, since the absolute number of AEA members located close to each potential convention site varies).

[4]Each AEA member in North America listed in the 1974 *Directory* was assigned to the major airport nearest to his or her home address. Airport locations were expressed in latitude and longitude readings from the *Times Atlas*. The few AEA members whose cities of reported residence could not be located in the *Rand McNally Road Atlas* were arbitrarily assigned to the capital city of their state of reported residence. The sample size is 16,909.

[5]Hans-Jürgen Zimmermann and Michael G. Sovereign (pp. 105–10) provide a thorough explanation of this method and a survey of some estimating techniques.

location). If the knot goes through a hole, costs are minimized at that hole's location.

II. Minimum AEA Convention Transportation Cost Estimates

Our particular problem is actually simpler than the distribution cost-minimization problem because there are only a limited number of potential locations for the AEA convention. Based upon available hotel space within walking distance of the convention center, the Secretary of the AEA identified twenty-six North American cities that might be possible sites for the annual convention. Consequently, we have evaluated the *TTC* equation for each of these locations. The total travel mileage to the site with the least *TTC* was assigned a value of one. Other sites' travel mileages were compared to it. The results are reported in Table 1, columns (1) and (2). This method minimizes miles traveled under the assumption of direct air routes from cities of origin to the convention location.

Pittsburgh, Pennsylvania is the location causing the fewest miles traveled to the convention, based on the 1974 distribution of AEA members.[6] To hold the convention in Cleveland instead of Pittsburgh under these same assumptions would raise total miles traveled by 1.9 percent. A San Francisco convention site would cause aggregate member miles traveled more than three times those of Pittsburgh. Table 1, column (1) can be used to quantify the relative miles traveled (and transportation costs if travel costs are linear with respect to distance) to alternative convention locations so that explicit decisions can be made weighing higher travel costs against lower food and/or lodging costs or greater benefits from cities other than Pittsburgh.

Average air fare per mile generally declines with stage length. In addition, variation in the extent of route competition and route costs (some routes of similar length incur different costs due to, for example, variations in weather) weakens the validity of the constant per mile cost assumption. In order to directly capture all of these effects we obtained the actual regular coach air fares in effect during September 1978 from each AEA member's home airport to each of the twenty-six potential convention sites.[7] Ground transportation

TABLE 1—RELATIVE TRAVEL COST[a] OF MEMBERSHIP TO VARIOUS POTENTIAL CONVENTION LOCATIONS

	Minimize Miles Traveled		Minimize Air Fares	
Location	Relative Cost (1)	Rank (2)	Relative Cost (3)	Rank (4)
Washington, D.C	1.066	5	1.000	1
Baltimore	1.028	3	1.000	2
Cleveland	1 019	2	1.019	3
Pittsburgh	1.000	1	1.022	4
Philadelphia	1 069	6	1.034	5
Detroit	1 048	4	1.036	6
New York	1 117	8	1.038	7
Chicago	1 131	9	1.104	8
Boston	1 326	13	1.193	9
St Louis	1 209	11	1.196	10
Atlanta	1 156	10	1.259	11
Toronto	1 081	7	1.271	12
Minneapolis	1.429	15	1 307	13
Kansas City	1 403	14	1.313	14
Montreal	1 309	12	1.387	15
New Orleans	1 451	16	1 492	16
Dallas	1 602	18	1.510	17
Houston	1 642	19	1.582	18
Miami	1 590	17	1.644	19
Denver	1 959	20	1.647	20
Seattle	3 166	25	2 065	21
Las Vegas	2 578	21	2.083	22
Los Angeles	2 759	23	2.201	23
San Diego	2 750	22	2.219	24
San Francisco	3 059	24	2 303	25
Honolulu	6 107	26	3.366	26

[a]All costs are expressed relative to the minimum cost location (1.000) The total air fare cost for the 16,909 AEA members if the convention were located in Washington, D C would be $2 673 million, based on September 1978 rates. Air fares for other locations can be computed by multiplying this figure by the ratio reported in col. (3).

[6]If the criteria of sufficient hotel space and air service are removed and equation (1) evaluated at each airport which is a home base for at least one economist, the mileage-minimization location is Charleston, West Virginia.

[7]The air fares were computed by *The Travel Agency*, Nashville TN. Published joint fares were used when available, constructed joint fares were used otherwise.

to and from home airports (which can vary substantially because some economists live substantial distances from the nearest airport) can be ignored because it does not vary by convention site. It was assumed that convention site transportation costs are the same for each city in order to simplify the calculations. Air fares from each home airport were weighted by the number of AEA members using the airport to derive an estimated total air fare bill for the twenty-six potential convention sites. The lowest air fare bill site was again assigned a value of one. The ratios of other sites' air fare bills to the lowest air fare bill are reported in Table 1, column (3). The rank order of potential convention sites minimizing September 1978 air fares is presented in Table 1, column (4).

The minimum transportation cost location using actual 1978 air fares is Washington D.C. Baltimore was only a few dollars more expensive. The location moves eastward from Pittsburgh because air fares generally decline with stage length and there are more AEA members in Washington D.C. than in Pittsburgh. Therefore the marginal cost of flying West Coast economists several hundred miles farther east and Pittsburgh economists to Washington D.C. is less than the savings available by eliminating the numerous high cost per mile trips of Washington D.C. economists traveling to Pittsburgh.

There are two major differences between the estimated minimum mileage location and the estimated minimum air fare location. Because of the declining fare per mile with increasing stage length, the West Coast locations, while continuing to be the most costly, are less disadvantageous in terms of actual air fares. For example, a San Francisco location would require about 305 percent more miles to be traveled vis-à-vis Pittsburgh, but only 230 percent greater air fare expenditures in contrast to Washington D.C. Second, because of higher per mile fares into Canada, Toronto and Montreal's air fare costs are relatively high in comparison with the minimum travel mileage necessary to hold conventions there. Toronto increases from a 8 to a 27 percent disadvantage as we compare mileage traveled to the air fare bill.

Although the air fare bill calculation minimizes out-of-pocket transportation costs, it does not account for all costs. For example, if the marginal discomfort of traveling increases with the length of the trip, the declining price per mile fare structure may understate total (monetary and nonmonetary) costs. The mileage-minimizing calculation reflects, at least to some extent, the possibly increasing marginal nonmonetary costs of travel with respect to distance. As Table 1 reveals, the cost-minimizing convention location moves westward the greater is the weight given to such nonmonetary costs.

III. Minimum Placement Meeting Transportation Cost Estimates

Conventioneers with less diverse goals than those who attend the AEA annual meeting may place an even greater emphasis on minimizing transportation costs. This is likely the case with special placement meetings such as those of January 1977 and January 1979, when the 1976 and 1978 AEA conventions were held earlier than the traditional after-Christmas period. Job candidates attending the placement meetings are usually intent on traveling to their destination, completing job interviews, and returning home to wait not so patiently by their telephones. Furthermore, most job candidates are personally responsible for their travel expenses.

We can easily apply the transportation cost-minimization model to the placement meetings. To do this, we computed the average annual number of new Ph.D.s produced by each institution during the 1972-74 period[8] and minimized their aggregate mileage traveled to a hypothetical placement meeting at alternative locations. We applied the assumptions that geographic distribution of attendance is proportional to the geographic distribution of degrees awarded, travel costs are constant with respect to distance, and travel is

[8]Number of Ph.D.s granted is taken from the 1974 *Guide to Graduate Study in Economics and Agricultural Economics.*

by straight line from home airport to the meeting location. We ignored the mileage traveled by potential employers. The projected air fare bill to job candidates for a placement meeting at each of the twenty-six cities was also calculated as the weighted (by number of Ph.D.s produced) sum of actual September 1978 air fares. The ranking and relative transportation costs of the twenty-six cities as placement locations are reported in Table 2.

TABLE 2—RELATIVE TRAVEL COST[a] OF NEW PH.D. ECONOMISTS TO VARIOUS POTENTIAL PLACEMENT MEETING LOCATIONS

Location	Minimize Miles Traveled		Minimize Air Fares	
	Relative Cost (1)	Rank (2)	Relative Cost (3)	Rank (4)
Chicago	1.000	1	1.000	1
Detroit	1.029	2	1.016	2
Cleveland	1.040	4	1.038	3
Pittsburgh	1.067	5	1.047	4
St. Louis	1.038	3	1.059	5
Baltimore	1.170	9	1.127	6[b]
Washington, D.C.	1.217	11	1.127	6[b]
Kansas City	1.153	8	1.138	8
Minneapolis	1.185	10	1.148	9
Philadelphia	1.220	12	1.151	10
New York	1.270	13	1.166	11
Atlanta	1.131	6	1.199	12
Boston	1.451	18	1.268	13
Toronto	1.132	7	1.306	14
Dallas	1.342	15	1.320	15
New Orleans	1.307	14	1.356	16
Denver	1.571	19	1.387	17
Houston	1.406	17	1.393	18
Montreal	1.401	16	1.524	19
Miami	1.584	20	1.606	20
Las Vegas	2.069	21	1.773	21
Los Angeles	2.251	23	1.872	22
San Diego	2.221	22	1.882	23
Seattle	2.558	25	1.926	24
San Francisco	2.458	24	1.940	25
Honolulu	5.115	26	2.931	26

[a]All costs are expressed relative to the minimum cost location (1.000). The total air fare cost for the annual average number of Ph.D.s if the placement meeting were held in Chicago would be $188,246, based on September 1978 air fares. Air fares for other locations can be computed by multiplying this figure by the ratio reported in col. (3).
[b]Tie.

Both mileage traveled and air fare expenditures by job candidates are minimized if the placement meeting is held in Chicago.[9] Consideration of air fares (vis-à-vis mileage) lessens the cost disadvantage of eastern cities with many meeting attendees, since locating the meeting in such cities eliminates numerous high per mile cost short trips. Boston and New York City, which graduate the most Ph.D.s, are only 26.8 and 16.6 percent more costly than Chicago on the basis of air fares, while they require 45.1 and 27.0 percent more travel mileage, respectively.

The Secretary of the AEA solicited site preferences for the 1979 placement meeting from placement officers at Ph.D.-granting institutions. Among Atlanta, Chicago, Kansas City, New Orleans, and St. Louis, the first choice was Chicago. Relative to Chicago, using the data in Table 2, column (3), the air fare bill disadvantages for St. Louis, Kansas City, Atlanta, and New Orleans are 5.9, 13.8, 19.9, and 35.6 percent, respectively. Would the job candidates be willing, as a group, to pay 35.6 percent more while traveling 30.7 percent more miles in order to hear jazz and be reminded that winter does not mean ice and snow everywhere while completing their job interviews?

IV. Conclusion

The minimum transportation cost AEA convention site is considerably east of the minimum transportation cost placement location (minimizing either mileage or air fares). This is because there is a larger proportion of AEA members in the East than in the Midwest and West than is the case for new economics Ph.D.s. The ten cities with the largest populations of AEA members and new Ph.D.s are reported in Table 3. New York, Washington, and Boston account for 28.1 percent of the AEA members but account for only 13.9 percent of new Ph.D.s. New Ph.D. economists seem to be centered similarly to

[9]If criteria for hotel accommodations and air service are disregarded, the mileage-minimizing location is Lafayette, Indiana.

TABLE 3—DISTRIBUTION OF AEA MEMBERS AND PH.D.S AMONG AIRPORTS SERVING TEN LARGEST POPULATIONS OF ECONOMISTS

Location	Percent of AEA Members, 1974	Location	Percent of New Ph.D.s, 1972–74
Washington, D.C.	14.1	Boston	6.1
New York	9.8	New York	5.4
Boston	4.2	Los Angeles	4.6
Los Angeles	4.1	San Francisco	4.6
Chicago	3.9	Syracuse	3.9
San Francisco	3.3	Champaign-Urbana	3.6
Baltimore	2.5	Chicago	3.4
Philadelphia	2.3	Madison, Wis	3.3
Newark	2.3	Raleigh-Durham	3.3
Toronto	1.5	Lafayette, Ind	2.8
		Lansing, Mich	2.8

the *U.S.* population in general,[10] but as they accept employment, they tend to gravitate eastward.

During the decade 1971–80 the AEA Convention was held (or is scheduled) in New Orleans, Toronto, New York, San Francisco, Dallas, Atlantic City, New York, Chicago, Atlanta, and Denver, respectively. The nine cities (excluding Atlantic City, which was not included in our twenty-six potential convention sites) average 51.9 percent above minimum mileage necessary and 46.3 percent above minimum air fares necessary for AEA conventions over the decade. Obviously something other than minimizing transportation costs (either mileage or air fares) is important in convention site selection.

Knowledge of relative transportation costs of AEA members (potential conventioneers) is not sufficient to set the site of future annual conventions, but it can usefully contribute to a more informed choice among alternative sites. At least the uncertainty and possibly erroneous choices emanating from distorted subjective evaluations of member transportation costs can be reduced by the data reported in this study. Significant elements of the choice among alternative convention sites remain unquantified (for example, relative lodging and meals costs, relative entertainment benefits, regional equity in the distribution of travel costs), but the decision will be improved if we have more precise information on one of its important components. With such information it is possible to better appreciate the cost of purchasing regional equity, lower lodging costs, greater entertainment benefits, etc. For example, if the only considerations in choosing between Washington and Chicago were transportation and lodging costs, it would be necessary to obtain a hotel rate approximately $5.48 dollars per person per night lower in Chicago than Washington to warrant a shift from the Capitol to the Windy City.[11] Such tradeoffs can be evaluated for each of the cities included in this assessment.

[10]The center of population for the entire United States (vis-à-vis North America) in 1975 was near St. Louis (see *Statistical Abstract of the United States: 1976*, p. 9).

[11]Round-trip transportation costs (based on air fare minimization) to Chicago are $16.44 per person greater than to Washington, D.C. Assuming a three-night hotel stay in either location, daily lodging costs would have to be at least $5.48 ($16.44/3) less in Chicago to make it the lowest travel *plus* lodging cost location, other things being equal. This calculation is based on an estimated difference in absolute air fare bills between Chicago and Washington, D.C. equal to $277,876 (approximately $2,673,000 x .104) for the 16,909 AEA members in the sample.

REFERENCES

Wyn F. Owen and Fred R. Glahe, *Guide to Graduate Study in Economics and Agricultural Economics*, 3d ed., Homewood 1974.

L. A. Thompson, R. M. Gaedeke, and D. H. Tootelian, "Professional Conferences—The Leisure of the Theory Class," *Collegiate News Views*, Winter 1977–78, *31*, 1–3.

Hans-Jürgen Zimmerman and Michael G. Sovereign, *Quantitative Models for Production Management*, Englewood Cliffs 1974.

American Economic Review, "1974 Directory of Members," Oct. 1974, suppl., *64*, No. 5.

Rand McNally Road Atlas, 52d ed., New York 1976.

Times Atlas of the World, Comprehensive Edition, Boston 1971.

U.S. Bureau of the Census, *Statistical Abstract of the United States: 1976*, Washington 1976.

Hedging and the Competitive Firm under Price Uncertainty

By DUNCAN M. HOLTHAUSEN*

Recent models of the competitive firm under uncertainty have explored ways in which the competitive firm's decisions would deviate from those of a firm operating with certainty and have explored the effects of increasing risk aversion on the firm's decisions (see, for example, Agnar Sandmo; David Baron; Hayne Leland; R. N. Batra and Aman Ullah; and the author). These models all assume that the firm's only response to uncertainty is to adjust its output or input levels. In practice, however, a number of institutions have been developed to aid firms in the management of risk.

One of the most common methods used to deal with price uncertainty is hedging. Futures markets exist for many agricultural products and some metals. Firms may also use forward contracts to fix the price at which output or inputs are traded in the future. Although there is a large literature dealing with futures markets, few attempts have been made to incorporate futures or forward trading in the theory of the firm.[1] Such a model is developed in this paper.

A risk-averse, competitive firm is assumed to face price uncertainty. It must choose its level of output before the price uncertainty is resolved and may, at the same time, buy or sell output in a forward market at a fixed price. The major result of the paper is that the firm will produce a level of output which depends only on the forward price and is, in particular, independent of the firm's degree of risk aversion and the probability distribution of the uncertain price. In addition, if the forward price is less than the expected future price, the firm will generally hedge some, but not all, output in the forward market; it will hedge more, the more risk averse it is; and it will hedge more as the riskiness of the uncertain price increases. Finally, the existence of a forward market will generally induce the firm to produce a greater output than it would have in the absence of such a market.

I. The Model

The competitive firm is assumed to face a stochastic market price p for its single output x. Output is produced at a cost of $c(x)$, with $dc/dx = c'(x) > 0$, and may either be sold in the future at the random market price or sold forward at the certain price b. The amount of output hedged in the forward market is given by h. The firm is assumed to have a von Neumann-Morgenstern utility function U defined on profit Π, and its goal is to maximize its expected utility of profit,

$$(1)\quad \max_{x,h} EU(\Pi) = \int_0^\infty U[p(x-h) + bh - c(x)]f(p)\,dp$$

where $f(p)$ is the firm's (subjective) probability density function on market price.

The first-order conditions[2] are

*Associate professor of economics and business, North Carolina State University

[1]See the papers by Ronald McKinnon, Leland Johnson, and Ronald Ward and Lehman Fletcher for some related models. After this paper was submitted for publication, a working paper by Gershon Feder, Richard Just, and Andrew Schmitz came to my attention. They have examined a model analogous to the one presented in this paper and have independently derived many similar results.

[2]Second-order conditions are assumed to hold for risk-averse firms. For risk-neutral or risk-loving firms, second-order conditions do not hold. To see this, take the second partial derivative with respect to h:

$$\frac{\partial^2 EU(\Pi)}{\partial h^2} = \int_0^\infty U''(\Pi)(b-p)^2 f(p)\,dp$$

This must be negative for an interior maximum. For the risk-averse firm, $U''(\Pi) < 0$ and the condition is satisfied. For a risk-neutral, $U''(\Pi) = 0$, or risk-loving firm, $U''(\Pi) > 0$, the second-order condition is violated. Thus, an interior maximum cannot be assumed for risk-neutral or risk-loving firms.

$$(2)\quad \frac{\partial EU(\Pi)}{\partial x} = \int_0^\infty U'(\Pi)[p - c'(x)] f(p)\, dp = 0$$

and

$$(3)\quad \frac{\partial EU(\Pi)}{\partial h} = \int_0^\infty U'(\Pi)(b - p) f(p)\, dp = 0$$

where $U'(\Pi) = dU(\Pi)/d\Pi$.

Condition (2) is exactly the same as that studied in detail by Baron and by Sandmo. Among other things, it can be shown that in the absence of forward markets, the risk-averse firm selects a lower level of output than the amount it would choose if market price were certain to be $E(p)$, and the amount of output decreases as aversion to risk increases.

The existence of a forward market modifies these results considerably. In particular, the firm's output level is independent of its degree of risk aversion. To see this, add conditions (2) and (3) to give

$$(4)\quad \int_0^\infty U'(\Pi)[b - c'(x)] f(p)\, dp = [b - c'(x)] \int_0^\infty U'(\Pi) f(p) d(p) = 0$$

The expression $[b - c'(x)]$ may be written outside the integral sign because neither b nor $c'(x)$ depends on the value of the uncertain price.

Since $U'(\Pi)$ is positive for any price, (4) holds only if $c'(x) = b$. The firm, therefore, chooses to produce that level of output for which marginal cost equals the certain forward price. Thus, all risk-averse firms in the market will key their production decisions to the forward price. Differences in risk aversion or in price expectations do not affect production decisions, although they do affect hedging decisions as will be shown later.

This result is interesting for two reasons. For one, if forward prices are widely known and fairly uniform across firms as is true in the commodity futures markets, then econometric supply models should use futures prices as the appropriate price variable.[3] Secondly, if forward markets are highly competitive, as commodity futures markets appear to be, we might expect b to be close to $E(p)$.[4] In this case, risk-averse firms will produce amounts close to the output levels that would be optimal if $E(p)$ were certain to occur. Therefore, price uncertainty and the existence of risk aversion would not have a significant effect on output.

The relation between the forward price b and the expected price $E(p)$ is important in the firm's hedging decision. To investigate this, rewrite (3) as follows,

$$(5)\quad E[U'(\Pi)(b - p)] = EU'(\Pi)\, E(b - p) + cov\, [U'(\Pi), -p] = 0$$

Expression (5) can be used to determine the conditions under which a risk-averse firm would hedge its entire output ($h = x$), hedge less than its entire output ($0 < h < x$), or speculate ($h > x$ and $h < 0$). The results can be summarized as follows. If the forward price equals the expected price ($b = E(p)$), the firm will hedge its entire output. If the forward price is less than the expected price, the firm will either hedge less than its entire output or, if the forward price is sufficiently less than the expected price, it will speculate by purchasing output in the forward market ($h < 0$) expecting to sell this plus its regular output at a greater price in the future. Finally, if the forward price is greater than the expected price, the firm will speculate by selling forward an amount greater than its output ($h > x$) expecting to purchase the additional output in the future at a low price in the open market.

These results depend on the sign of the covariance term in (5). For a risk-averse firm the covariance is zero (positive)(negative) if and only if $h = x (h < x)(h > x)$. To see this recall that the firm's profit is given by $\Pi = p(x - h) + bh - c(x)$. If $h = x$, profit is not

[3]See Bruce Gardner for an example of how well futures prices work in a supply model.

[4]A number of authors have tested for bias in futures market prices. Most have not found significant biases. Some examples of this literature are Charles Rockwell, William Tomek and Roger Gray, and Gray.

a function of price, and hence the covariance is zero. If $h < x$, price is multiplied by a positive value in the profit equation, and profit is directly related to price. Therefore, increases in price increase profit and reduce $U'(\Pi)$. Since increases in price also reduce $-p$, the covariance is positive. Finally, if $h > x$, price is multiplied by a negative value in the profit equation, and following the same logic as before, covariance is negative. Now, if $b = E(p)$, the covariance must be zero for (5) to hold. Hence $h = x$. Similarly if $b < E(p)$, covariance must be positive which implies $h < x$, and if $b > E(p)$, covariance must be negative implying $h > x$.

In the traditional Keynesian view of hedging, the forward price is less than the expected price. In this case the firm hedges some but not all of its output (unless the forward price is so low that the firm is induced to speculate as noted above). For the units hedged, the firm is willing to accept a forward price less than the expected price in order to avoid the price uncertainty on those units. The difference between the expected price and the forward price is an insurance premium the hedger pays to the individual (usually described as a speculator) who is on the other end of the transaction.

II. The Effect of Increasing Risk Aversion

Since the level of output x is independent of the degree of risk aversion, we can determine the effect of increasing risk aversion by examining condition (3). Let the degree of risk aversion be measured by the Pratt-Arrow index, $r(\Pi) = -U''(\Pi)/U'(\Pi)$, where ($'$) and ($''$) stand for first and second derivatives, respectively. Defining the concept of "increasing risk aversion" to mean that the firm would be willing to pay an increasing amount to insure against a given risk, Pratt has shown that the index $r(\Pi)$ increases as the utility function becomes more risk averse.

Using the Pratt-Arrow index, the following results hold. If the forward price is less than the expected price, the firm will hedge a greater amount of output as it becomes more risk averse. If the forward price is greater than the expected price, so that the firm speculates (i.e., $h > x$ as shown earlier), the more risk averse it becomes the less it will speculate. These results are proved in the Appendix.

The conclusions are intuitively appealing. We expect risk averters to be willing to trade off higher expected profits for lower certain profits. When the forward price is less than the expected price, the firm does give up some expected profit as it increases its hedge. Similarly, when the forward price is greater than the expected price, the firm gives up some expected profit when it reduces its "hedge." Since the hedge in this case is actually a speculative forward sale, the result is what we would expect.

III. The Effect of Increasing Risk

Following Michael Rothschild and Joseph Stiglitz, increasing risk is defined as a mean preserving spread of the distribution of prices. A mean preserving spread shifts some weight from the center to the tails of the distribution leaving the mean of the distribution constant. Define p^* as $p^* = ap + m$ where a and m are shift parameters initially set equal to 1 and 0, respectively. A mean preserving spread of the density function of p^* requires that the expected price remain constant, or $dE(p^*) = dE(ap + m) = \mu da + dm = 0$, where $E(p) = \mu$. Therefore, $dm/da = -\mu$.

We already know that the firm's production decision is not influenced by the distribution of p since $c'(x) = b$. Therefore, increasing risk will only influence h. To examine this influence, replace p with p^* in (3). Profits are now given by $\Pi = p^*(x - h) + bh - c(x) = (ap + m)(x - h) + bh - c(x)$. The first-order condition can now be written as

$$(6) \qquad E[U'(\Pi)(b - ap - m)] = 0$$

Differentiating (6) with respect to a and h, and using $dm/da = -\mu$ gives

$$(7) \quad \frac{dh}{da} = -\{E[U''(\Pi)(p - \mu)(x - h) \cdot (b - ap - m) + U'(\Pi)(\mu - p)]\} \div [EU''(\Pi)(b - ap - m)^2]$$

The denominator of (7) is negative by second-order conditions, and thus, the sign of dh/da is the same as the sign of the numerator.

The sign of the curly bracketed term in the numerator of (7) can be shown to be positive if $h < x$ and the firm is constant or decreasingly absolute risk averse. Proof is given in the Appendix. Kenneth Arrow, among others, has argued that individuals are characterized by decreasing absolute risk aversion. If this is also a reasonable assumption for firms, then $dh/da > 0$ when $h < x$. This implies that the amount hedged would be increased as the riskiness of the probability distribution on price increases. If $h > x$ the result is ambiguous.

IV. Effects on Output

A question of some importance is whether the existence of forward markets will induce the risk-averse firm to produce more or less output than it would without those markets. To explore this question, think of h for the moment as a parameter out of the firm's control. Denoting $EU(\Pi)$ as ϕ, take the total differential of the first-order condition (2) with respect to x and h. This yields

$$\frac{dx}{dh} = -\frac{\phi_{xh}}{\phi_{xx}} \tag{8}$$

where subscripts denote partial derivatives. We know $\phi_{xx} < 0$ from second-order conditions, so the sign of dx/dh is the same as the sign of ϕ_{xh}. Writing out this cross-partial term gives

$$\phi_{xh} = \int_0^{\infty} U''(\Pi) \cdot [p - c'(x)](b - p) f(p)\, dp \tag{9}$$

For a risk-averse firm, $U''(\Pi) < 0$. At the optimal hedge, $c'(x) = b$, and so $[p - c'(x)]$ and $(b - p)$ will be of opposite sign for every value of p. Thus, $[p - c'(x)](b - p)$ is negative, and $\phi_{xh} > 0$. It follows that $dx/dh > 0$.

Hence, if h were originally set at the optimal value and then reduced below that value, the firm would reduce its output. Thus a firm that is restricted to a hedge less than the optimal amount responds by decreasing output. Technically, this result only holds in a neighborhood of the optimal hedge. The same result can be shown to hold globally, however, by maximizing (1) subject to the constraint $h = \bar{h}$ where $\bar{h}$ is any value below the optimal hedge. The existence of forward or futures markets therefore induces a firm to increase output, *ceteris paribus*, assuming its optimal hedge is positive.[5] If the optimal hedge is negative, output would be reduced when hedging becomes available.

V. An Example

To illustrate the analysis presented in this paper, consider the following example. Let $f(p)$ be normal with mean μ and variance σ^2. Let utility be given by the constant risk averse function $U(\Pi) = -e^{-\alpha\Pi}$, where $\alpha > 0$. Then the measure of risk aversion is $r(\Pi) = \alpha$. It is well known that maximization of expected utility in this case is equivalent to maximization of $\mu_\Pi - \alpha\sigma_\Pi^2/2$, where μ_Π and σ_Π^2 are the mean and variance of profits (see, for example, Baron, p. 469). Substituting the appropriate expressions for μ_Π and σ_Π^2 gives $\mu(x - h) + bh - c(x) - \alpha(x - h)^2\sigma^2/2$ to be maximized. Taking both first-order conditions and solving simultaneously yields $c'(x) = b$ as required. From the first-order condition on h, the optimal values of x and h must satisfy

$$x - h = \frac{\mu - b}{\alpha\sigma^2} \tag{10}$$

Therefore, in the Keynesian case where μ exceeds b, x is greater than h, and some output will not be hedged. If $\mu = b$, then $x = h$, and all output is hedged. And if $\mu < b$, then $x < h$, and the firm speculates by selling forward an amount greater than output.

If the firm's degree of risk aversion α increases, $(x - h)$ decreases. Since x is independent of α, it follows that the amount hedged rises as α increases. If σ^2 increases, the riskiness of the distribution increases and the amount hedged increases.

Finally, we can see what condition would lead a firm to choose $h < 0$. Baron has shown

[5] A similar statement for the total output of all firms in a market would require a general equilibrium model to consider the effect of greater output on both the market price and the forward price.

that in the absence of a forward market, the firm would produce an amount such that $c'(x) = \mu - \alpha x\sigma^2$. Now, if the forward price b were less than $\mu - \alpha x\sigma^2$, the firm's output would satisfy $c'(x) = b < \mu - \alpha x\sigma^2$, and the firm would produce less than it would have in the absence of the forward market. Substituting $\mu - \alpha x\sigma^2$ for b in (10) yields

$$x - h < \frac{\mu - (\mu - \alpha x\sigma^2)}{\alpha\sigma^2} = x$$

which implies $h < 0$. Thus, only if the forward price is less than the certainty equivalent price $\mu - \alpha x\sigma^2$ would a firm desire to purchase output in the forward market.

VI. Summary

A model of the competitive firm under price uncertainty has been analyzed to show the effect that futures or forward markets would have on the firm's decisions. The existence of such markets is shown to be very important, since production decisions will be based on the forward price and not on the price the firm expects to prevail. In addition, the degree of risk aversion will have no influence on the production decision. Risk aversion affects the firm's optimal hedge, and if the forward price is less than the expected price, the hedge increases as the firm becomes more risk averse. Also, if firms are characterized by nonincreasing absolute risk aversion, the optimal hedge increases as the riskiness of the price uncertainty increases.

These results clearly show that the firm behaves quite differently when forward markets exist. Further work in this area ought to consider two aspects omitted from this analysis. One is to incorporate both production and price uncertainty, and the other is to put the analysis in a general equilibrium setting so that questions regarding aggregate output and market prices can be addressed.

APPENDIX

PROOF *that the optimal hedge increases (decreases) as the firm becomes more risk averse given $h < x$ ($h > x$)*:

Consider two firms with concave utility functions $U_1(\Pi)$ and $U_2(\Pi)$ with $r_1(\Pi) > r_2(\Pi)$ for all Π, where r_1 and r_2 are the Pratt-Arrow indexes of risk aversion for U_1 and U_2, respectively. For firm 1, first-order condition (3) may be rewritten as

$$\text{(A1)} \qquad \int_0^\infty \frac{U_1'(\Pi)}{U_1'(\Pi_0)}(b-p)f(p)\,dp = 0$$

where Π_0 is a constant defined below, and the condition is satisfied at the optimal hedging level for firm 1, h_1^*. Define p_0 to be the price which equals the forward price b, and let Π_0 be the firm's profit when $p = p_0$. Then (A1) can be expressed as the sum of two integrals:

$$\text{(A2)} \qquad \int_0^{p_0} \frac{U_1'(\Pi)}{U_1'(\Pi_0)}(b-p)f(p)\,dp + \int_{p_0}^\infty \frac{U_1'(\Pi)}{U_1'(\Pi_0)}(b-p)f(p)\,dp = 0$$

Now write the first-order condition for firm 2 evaluated at h_1^* as

$$\text{(A3)} \qquad \int_0^{p_0} \frac{U_2'(\Pi)}{U_2'(\Pi_0)}(b-p)f(p)\,dp + \int_{p_0}^\infty \frac{U_2'(\Pi)}{U_2'(\Pi_0)}(b-p)f(p)\,dp$$

Subtracting (A2) from (A3) yields

$$\text{(A4)} \qquad \int_0^{p_0} \left[\frac{U_2'(\Pi)}{U_2'(\Pi_0)} - \frac{U_1'(\Pi)}{U_1'(\Pi_0)}\right] \cdot (b-p)f(p)\,dp + \int_{p_0}^\infty \left[\frac{U_2'(\Pi)}{U_2'(\Pi_0)} - \frac{U_1'(\Pi)}{U_1'(\Pi_0)}\right] \cdot (b-p)f(p)\,dp$$

We want to find the sign of (A4). Assuming the optimal hedge is less than output, which would be the case when the forward price is less than the expected price, profit increases with increases in price, and therefore $\Pi_0 > \Pi$ in the first integral and $\Pi_0 < \Pi$ in the second integral. Then, using John Pratt's equation (20), the term in brackets in the first integral is negative and the similar term in the second integral is positive. Also, by definition of p_0, $(b - p)$ is positive in the first integral and negative in the second integral. Therefore, each integral is negative, and (A4) is negative. This implies that (A3) is negative, and in

order to drive (A3) to zero, the amount hedged must be reduced from h_1^*. Hence $h_2^* < h_1^*$; the more risk-averse firm (firm 1) hedges a greater amount of output. So as risk aversion increases, the size of the optimal hedge also increases.

If the forward price is greater than the expected price, so that the optimal hedge exceeds output, the signs of the bracketed terms in (A4) are reversed. Therefore (A4) is positive which implies $h_2^* > h_1^*$. The more risk-averse firm hedges a smaller amount of output. In this case then, as risk aversion increases the size of the optimal hedge decreases.

PROOF *that the numerator of equation (7) is positive if $h < x$ and U is nonincreasingly absolute risk averse*:

Since we are interested in the sign of dh/da as the distribution of prices changes from its current shape, we need to evaluate (7) at the point $a = 1$ and $m = 0$. Thus, the sign of (7) is the same as the sign of

$$\text{(A5)}\quad E(U''(\Pi)(p-\mu)(x-h)(b-p) + U'(\Pi)(\mu-p)] = (x-h)E[U''(\Pi)(p-\mu)(b-p)] + E[U'(\Pi)(\mu-p)]$$

The second term in (A5) can be shown to be positive if $h < x$. Write that term as

$$\text{(A6)}\quad E[U'(\Pi)]E(\mu-p) + cov\,(U'(\Pi), -p)$$

Since $E(\mu - p) = 0$, and since $cov\,(U'(\Pi), -p)$ was shown to be positive in Section I, (A6) is positive.

To evaluate the first term on the right-hand side of (A5) substitute $(b-\mu) - (b-p)$ for $(p-\mu)$ and write the term as

$$\text{(A7)}\quad -E[U''(\Pi)(b-p)^2] + (b-\mu)E[U''(\Pi)(b-p)]$$

The first term in (A7) is positive since $U''(\Pi) < 0$. If $h < x$, then $(b - \mu)$ is negative from Section I. It remains only to find the sign of $E[U''(\Pi)(b-p)]$. As before, define $p_0 = b$ and Π_0 as profit when $p = p_0$. Then for nonincreasing absolute risk aversion and $h < x$,

$$\text{(A8)}\quad r(\Pi) \leq (\geq)\ r(\Pi_0) \qquad \text{for } p \geq (\leq)\ p_0$$

Substituting the definition for $r(\Pi)$ gives

$$\text{(A9)}\quad -U''(\Pi) \leq (\geq)\ r(\Pi_0)U'(\Pi) \qquad \text{for } p \geq (\leq)\ p_0$$

Now multiply (A9) by $(p_0 - p)$, take expectations of both sides, and use $p_0 = b$ to yield

$$\text{(A10)}\quad -E[U''(\Pi)(b-p)] \geq r(\Pi_0)E[U'(\Pi)(b-p)] \quad \text{for all } p$$

Equation (A10) holds for all p because the inequality is reversed when (A9) is multiplied by $(p_0 - p)$ and $p \geq p_0$.

The right-hand side of (A10) is zero by the first-order condition (3). Therefore, $E[U''(\Pi)(b-p)] \geq 0$, and (A7) is thus positive. This in turn implies that (A5) is positive and finally that (7) is positive. Hence, $dh/da > 0$ as was to be shown. Note that if $h > x$, (A7) can still be shown to be positive, but (A6) will be negative and the result is then ambiguous.

REFERENCES

Kenneth J. Arrow, *Essays in the Theory of Risk Bearing*, Chicago 1971.

D. P. Baron, "Price Uncertainty, Utility, and Industry Equilibrium in Pure Competition," *Int. Econ. Rev.*, Oct. 1970, *11*, 463–80.

R. N. Batra and A. Ullah, "Competitive Firm and the Theory of Input Demand under Price Uncertainty," *J. Polit. Econ.*, May/June 1974, *82*, 537–48.

G. Feder, R. E. Just, and A. Schmitz, "Futures Markets and the Theory of the Firm Under Price Uncertainty," work. paper no. 52, Giannini Found. Agr. Econ., Univ. California, Dec. 1977.

B. L. Gardner, "Futures Prices in Supply Analysis," *Amer. J. Agr. Econ.*, Feb. 1976, *58*, 81–84.

R. W. Gray, "The Search for a Risk Premium," *J. Polit. Econ.*, June 1961, *69*, 250–60.

D. M. Holthausen, "Input Choices and Uncertain Demand," *Amer. Econ. Rev.*, Mar. 1976, *66*, 94–103.

L. L. Johnson, "The Theory of Hedging and Speculation in Commodity Futures," *Rev. Econ. Stud.*, June 1960, *27*, 139–51.

H. E. Leland, "Theory of the Firm Facing Uncertain Demand," *Amer. Econ. Rev.*, June 1972, *62*, 278–91.

R. I. McKinnon, "Futures Markets, Buffer Stocks, and Income Stability for Primary Producers," *J. Polit. Econ.*, Dec. 1967, *75*, 844–61.

J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, Jan./Apr. 1964, *32*, 122–36.

C. S. Rockwell, "Normal Backwardation, Forecasting, and the Returns to Commodity Futures Traders," *Food Res. Inst. Stud.*, 1967, Suppl., *7*, 107–30.

M. Rothschild and J. E. Stiglitz, "Increasing Risk I: A Definition," *J. Econ. Theory*, Sept. 1970, *2*, 225–43; "II: Its Economic Consequences," Mar. 1971, *3*, 66–84.

A. Sandmo, "On the Theory of the Competitive Firm under Price Uncertainty," *Amer. Econ. Rev.*, Mar. 1971, *61*, 65–73.

W. G. Tomek and R. W. Gray, "Temporal Relationships Among Prices on Commodity Futures Markets: Their Allocative and Stabilizing Roles," *Amer. J. Agr. Econ.*, Aug. 1970, *52*, 372–80.

R. W. Ward and L. B. Fletcher, "A Micro Model of Optimal Futures and Cash Market Positions," *Amer. J. Agr. Econ.*, Feb. 1971, *53*, 71–78.

Unanticipated Money Growth and Unemployment in the United States: Comment

By DAVID H. SMALL*

In recent years, the natural rate of unemployment hypothesis and the corresponding long-run neutrality of monetary and fiscal policy have been accepted by a growing proportion of the economics profession. Robert Lucas, and Thomas Sargent and Neil Wallace have used the theory of rational expectations to extend this neutrality to the short run. If valid, this extension is a death blow to short-run stabilization policy. Until recently, these short-run models have received little empirical support. Robert Barro's recent paper in this *Review* claims to rectify this shortcoming.

In implicit form, Barro's formulation of the short-run rational expectations model is equation (1) below, which states that "only unanticipated movements in money affect real economic variables" (p. 101) like the unemployment rate.

$$U = U^* + f(DM - DM^e) \tag{1}$$

where $f'(\ )$ is negative, U = the unemployment rate, U^* = the natural unemployment rate, DM = the rate of growth of the money supply, and DM^e = the rational expectation of DM. The principal obstacle to an empirical test of equation (1) is the absence of information on the expectations of economic agents. Barro's strategy is to specify a money supply model and equate the model's prediction errors with unanticipated monetary movements ($DM - DM^e$). When this measure of monetary surprises is inserted into an estimable form of (1), the equation as a whole and the monetary surprise variables individually are found to be statistically significant. As a result, Barro claims to have demonstrated that the rational expectations model of unemployment is supported by the data.

I will argue that significant support for equation (1) has not been demonstrated. In Section I, Barro's theoretical model of the model supply is accepted, but his empirical implementation of the theory is shown to be flawed. When agents are allowed to be a bit more rational and intelligent in forming expectations about monetary behavior, the empirical support for equation (1) is severely weakened. Deviations in unemployment from the natural rate are shown to be consistent with standard macro-economic theory in which monetary and fiscal policy are effective in the short run. In Section II, Barro's measure of the natural rate of unemployment in equation (1) is rejected for purely empirical reasons.

I. Analysis of Monetary Surprises

In this section Barro's measurement of unanticipated monetary movements is questioned. Is it possible that the money equation leaves as prediction errors monetary movements which actually could have been predicted in advance?

Equation (2) is Barro's empirical money supply model, with standard errors in parentheses, and where $\hat{\sigma}$ is the standard error of the regression:

$$\begin{aligned} DM_t = \underset{(.031)}{.087} &+ \underset{(.15)}{0.24}DM_{t-1} \\ &+ \underset{(.13)}{0.35}DM_{t-2} + \underset{(.015)}{.082}FEDV_t \\ &+ \underset{(.010)}{.027}UN_{t-1} \end{aligned} \tag{2}$$

$$R^2 = .90,\ D.W. = 2.39,\ \hat{\sigma} = .020$$

where DM is the annual rate of change in the money supply (M_1); UN is an unemployment

*Graduate student, Northwestern University. I would like to thank Robert M. Coen, Robert J Gordon, Dale T Mortensen, Christopher Sims, and James A. Wilcox for helpful comments, Robert Barro for making his data base available, and the Social Science Research Council-Subcommittee on Monetary Research for financial assistance.

variable designed to pick up countercyclical reactions of the monetary authority; and *FEDV*, a measure of actual federal expenditures in excess of normal expenditures, captures money creation required by federal financing needs.

The basis of Barro's theory of federal finance is a "fixed capital [stock] that has been accumulated in the tax-raising capacity" (p. 101). An excess of federal expenditures over their "normal" level strains the tax-raising capacity of the capital stock. This strain is reduced by increasing the supply of capital (and thus increasing taxation) and by reducing the demand on the stock by financing part of the increase in expenditures by increasing money creation. For a given increase in expenditures, the split between increased taxation and money creation will depend on whether the increase is permanent or temporary, and whether it is anticipated or unanticipated.

The present value of the opportunity cost of not investing in tax-raising capital (and thus of not raising taxes in response to higher expenditures) is smaller when the increase in expenditures is temporary rather than permanent. Assuming investment is subject to increasing marginal adjustment costs, increased taxation is less attractive and money creation more attractive when a given increase in expenditures is temporary rather than permanent. Although Barro noted the importance of applying the permanent-temporary distinction to an increase in expenditures, the distinction is not embodied in equation (2) (see Barro, p. 103). To see this, assume that the expenditure changes shown in Figure 1 are foreseen with certainty. The movement of federal expenditures over intervals *A* and *B* are identical. Since *FEDV* depends only on present and past values of federal expenditures, *FEDV* and the incentive for money creation in equation (2) are the same over the two intervals.[1] However, money creation should be greater over the given time interval in the case of the temporary increase, as shown in the left-hand diagram. This

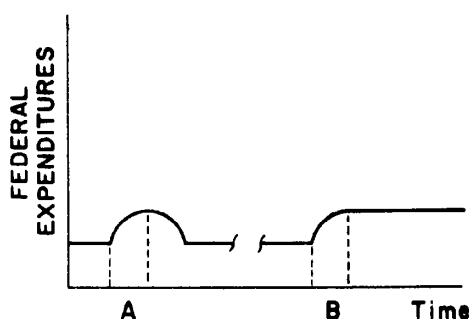



FIGURE 1

argument implies that the coefficient on *FEDV* should be greater when the increase in federal expenditures being measured by *FEDV* is temporary. The failure to allow this coefficient to vary may result in a systematic underprediction of the rate of growth of the money supply during periods of temporary increases in federal expenditures.

Since Barro's sample period includes "periods of wartime during which there are sharp temporary increases in the government budget" (p. 103), the above analysis, which could have been carried out by any rational agent in 1946, predicts that Barro's unanticipated money growth rates will have a positive bias during war years and (since the residuals must sum to zero) a negative bias during nonwar years. The following stylized facts about the money surprises indicate the predicted structure.

(a) Ranking the residuals from (2) in decreasing order yields the yearly ordering: 1968, 1969, 1950, 1951, 1971, 1946, 1952, ...[2] The residuals during all these years are positive. The Korean and Vietnam war years dominate this ranking.

(b) Over the nonwar interval between 1956 and 1963 seven out of eight money residuals are negative, whereas over the interval from 1964 to 1969 five out of six are positive.

This structure to the measures of unanticipated monetary movements is inconsistent with rational expectations since, on the basis

[1] $FEDV = log(FED) - [log(FED)]^*$, where $[log(FED)]_t^* = \beta[log(FED)]_t + (1-\beta)[log(FED)]_{t-1}^*$, and FED = real federal expenditures.

[2] This ranking does not include the residuals prior to 1964 since they were not used in Barro's unemployment equation.

of the previous theoretical analysis, it could have been predicted prior to the formation of the expectations.

A formal test of the variability of the *FEDV* coefficient in equation (2) can be developed by adding to equation (2) a new variable (*DFEDV*) which equals *FEDV* times a dummy variable. The dummy is set equal to one during war years and to zero otherwise.[3] The coefficient of *DFEDV* measures the increased impact of *FEDV* during years of temporary increases in expenditures. The null hypothesis that the coefficient of *DFEDV* equals zero corresponds to Barro's original formulation of the money supply process. The alternative hypothesis is that this coefficient is greater than zero, that is, temporary increases in expenditures cause more money creation than permanent ones. The estimated coefficients of the new money equation are

$$(3)\quad DM_t = \underset{(.027)}{.10} + \underset{(.169)}{.40DM_{t-1}} + \underset{(.14)}{.31DM_{t-2}} + \underset{(.013)}{.07FEDV_t} + \underset{(.013)}{.07DFEDV_t} + \underset{(.009)}{.03UN_{t-1}}$$

$$R^2 = .91,\ \hat{\rho} = \underset{(.161)}{-.42},\ \hat{\sigma} = .019$$

where $\hat{\rho}$ = the autocorrelation coefficient estimated by the Cochrane-Orcutt technique, and standard errors are reported in parentheses. Barro's restriction is rejected since the coefficient on *DFEDV* is significantly positive. Thus, Barro's restriction caused a structure to be present in his residuals.[4] Since agents are clearly aware of the existence of war, this data belongs in the money supply equation, as in (3). The residuals from equation (3) are used as the measure of unanticipated monetary movements in equation (5). Equation (4) is Barro's estimated unemployment equation:

$$(4)\quad \log U/(1-U) = \underset{(.15)}{-3.07} - \underset{(2.1)}{5.8DMR_t} - \underset{(1.9)}{12.1DMR_{t-1}} - \underset{(1.9)}{4.2DMR_{t-2}} - \underset{(0.8)}{4.7MIL_t} - \underset{(.46)}{0.95MINW_t}$$

$$R^2 = .78,\ D.W. = 1.96,\ \hat{\sigma} = .13$$

where *DMR* = the residuals from equation (2), *U* = the unemployment rate, *MIL* = a military draft variable, and *MINW* = a minimum wage variable. The latter two variables are determinants of the natural rate of unemployment.

$$(5)\quad \log U/(1-U) = \underset{(.20)}{-2.69} - \underset{(3.37)}{4.55DMS_t} - \underset{(3.01)}{10.60DMS_{t-1}} - \underset{(3.34)}{4.64DMS_{t-2}} - \underset{(1.16)}{4.55MIL_t} - \underset{(.60)}{.35MINW_t}$$

$$R^2 = .53,\ D.W. = 1.04,\ \hat{\sigma} = .19$$

[3]The dummy variable for *DFEDV* was set equal to one for the years [1950–52, 1965–70]. Other sets of years tried were [1950–53, 1965–70], [1950–53, 1968–70], [1951–52, 1965–70], [1951–52, 1968–70], [1951–53, 1965–70], [1951–53, 1968–70], [1950–52, 1968–70] and [1950–52, 1967–70] The *t*-statistic varied from 1 61 to 2.60 with an average of 2.1. In determining the correct dates for the dummy variables it must first be determined when the wars started and ended The Korean War clearly started in June of 1950 but the end could be considered the end of the fighting in late 1952, the signing of the truce in June of 1953, or even later due to the continued *U.S.* presence in South Korea Similar problems exist with respect to the Vietnam War. Additionally, the lag with which the *U.S* citizens learned of the war must be determined If the lag was one year then the dummy variable for 1950 should equal zero It should equal one if there was no information lag. In the text no information lag was assumed. Dummy variables were not included for World War II because it was most likely perceived as quite different than the Korean and Vietnam wars. A dummy variable for World War II can be included with no significant effects on the results if the new dummy is entered independently of the dummy for the Korean and Vietnam wars, i e., as long as the agents are not forced to believe all wars have the same impact on the money supply. As Barro notes, expectations for, say, 1950 based on an equation fitted through 1973 implies information was being used that was not available at the time. Preferably a rolling sample method should be used in which the regression is run only through the year prior to the year whose rate of monetary growth is being predicted. This method was not used for two reasons First, I wanted to stay as close as possible to Barro's reported technique. Second, using this method in concurrence with dummy variables would imply a perfect fit for the first year the dummy was equal to one

[4]Otherwise the coefficient of *DFEDV* would be insignificant The ranking of the residuals from equation (3) in decreasing order starts with the ordering [1950, 1951, 1971, 1969, 1972, 1973, 1959, 1947, 1955, . . .]. The war years do not dominate this ordering as much as they dominated the ordering of the residuals from equation (2)

where DMS = the residuals from equation (3) and all other variables are defined above.

The statistical significance of the unemployment equation as a whole and of the money surprises individually fall dramatically as the residuals from equation (3) are substituted for those from (2). The positive bias in Barro's monetary surprises during periods of temporary increases in federal expenditures accounts for his ability to explain the large decreases in unemployment during these intervals. Over the years (1951–53, 1966–69) the sum of the unemployment residuals (actual minus predicted) from equation (4) is −.2478 and from equation (5) is −.93084.[5] This is an increase of 276 percent. Thus, removal of the bias in the residuals from equation (2) decreases the unemployment equation's ability to track the large decreases in unemployment that occurred during these years.

The preceding analysis considered the differential impact of temporary vs. permanent changes in federal expenditures on money creation. It also matters whether a given change in expenditures is anticipated or unanticipated. If an increase in expenditures is anticipated, investment in tax-raising capital is possible prior to the increase in expenditures. Assuming increasing marginal adjustment costs, the average cost per unit of investment can be held down by spreading the investment over a longer time period when increases in expenditures are anticipated rather than unanticipated. Thus, more investment and less money creation are optimal when an increase in expenditures is anticipated as opposed to unanticipated.

The residuals from equation (6) are measures of unanticipated increases in federal expenditures. Divided by the money stock of the previous period, they measure the incentives for increases in the rate of growth of the money supply due to unanticipated increases in expenditures. This new variable is called $FEDUN$. Nominal GNP enters equation (6) because, for a given tax structure, government revenue increases and thus the budget constraint is relaxed with growth in GNP.

$$(6)\quad FED_t = \underset{(3.3)}{2.6} + \underset{(.14)}{1.2FED_{t-1}} - \underset{(.13)}{.67FED_{t-2}} + \underset{(.019)}{.08GNP_{t-1}},$$

$$R^2 = .98,\ D.W. = 1.98$$

where FED = nominal federal outlays and GNP = nominal GNP.[6]

The following new model of the money supply process not only takes into account the distinction between anticipated and unanticipated expenditure changes, but also quantifies the impact of war on the money supply. A measure of the deviation in the number of military personnel from "normal" was derived for each year of World War II, the Korean War, and the Vietnam War. These deviations were multiplied by a cost per man-year figure for each war to yield a measure of the cost of the war per year. These costs were divided by the money stock. The term $MILX1$ includes the cost per year of World War II and is zero otherwise. The Korean and Vietnam War costs are in the $MILX2$ series.[7]

$$(7)\quad DM_t = \underset{(.050)}{.25} + \underset{(.12)}{.47DM_{t-1}} + \underset{(.12)}{.11DM_{t-2}} + \underset{(.018)}{.05MILX1_{t-1}} + \underset{(.097)}{.34MILX2_{t-1}} + \underset{(.03)}{.21FEDUN_t} + \underset{(.017)}{.079UN_{t-1}}$$

$$R^2 = .94,\ D.W. = 2.00,\ \hat{\sigma} = .017$$

Equation (7) fits the data better than equation (2).[8] The residuals from equation (7),

[5]The average value of $log\ U/(1 - U)$ over these years is −3.4.

[6]Equation (6) can be run in real terms with no significant effect on the results: FED, through 1970, from *Historical Statistics of the United States, Colonial Times to 1970*; 1971–73 from *Statistical Abstract of the United States, 1975*, p. 226; GNP, through 1946, from *Historical Statistics of the United States*, p. 224; 1947–73 from *Survey of Current Business*, p. 50.

[7]See the Appendix for a more complete description of the derivation of $MILX1$ and $MILX2$.

[8]In equations (2) and (7) current period federal expenditures are present through the variables $FEDV$

called DMT, are used as the measure of unanticipated monetary movements in equation (8).

$$(8)\quad \log U/(1-U) = \underset{(.26)}{-2.73} + \underset{(3.24)}{4.25DMT_t} - \underset{(3.21)}{1.49DMT_{t-1}} - \underset{(3.12)}{4.16DMT_{t-2}} - \underset{(1.39)}{2.96MIL_t} - \underset{(.75)}{.53MINW_t}$$

$$R^2 = .31,\ D.W. = .82,\ \hat{\sigma} = .23$$

With this new set of monetary surprises, Barro's unemployment equation becomes statistically insignificant. Equation (8)'s F-statistic of 1.96 is less than the 5 percent critical value. The sum of the residuals from equation (8) over the years 1951–53, 1966–69 is −1.36 compared to −.2478 for the unemployment residuals from equation (4); an increase of 447 percent. Thus again, when the bias is removed from Barro's monetary surprise variable the resulting unemployment equation is unable to explain the large decreases in unemployment during war years.

The analysis and empirical results of this section indicate that Barro used a poor money equation which artificially excluded information available to the economic agents at the time. Although he recognized both that permanent and temporary increases in federal expenditures will have a different impact on money creation and that war results in large temporary increases in federal expenditures, these factors were not included in the agents' information set. The result was that the money equation consistently underpredicted the major wartime increases in the rate of growth of the money supply. This positive bias in his money surprises contains information on these major increases in money and is crucial to Barro's ability to explain the large wartime swings in the unemployment rate.

Thus, if anything, Barro has provided evidence that anticipated changes in monetary policy affect unemployment in the short run.[9]

II. Analysis of the Natural Rate of Unemployment

This section questions Barro's measurement of the natural rate of unemployment which is used to explain the increase in unemployment in the 1970's.

The unemployment rate equaled 3.4 percent in 1969, jumped to 4.7 percent in 1970, and remained high in 1971 and 1972. An increase in the estimated natural rate of unemployment due to a change in the military draft variable (MIL) seems to be a major reason for the ability of Barro's unemployment equation to track the increase in the actual unemployment rate.[10] The change in MIL was a result of Barro setting the variable equal to zero in 1970. This induced the largest change in both MIL and in the estimated natural unemployment rate over the entire sample period. This rather drastic change in MIL may reflect real economic events or may be an *ad hoc* change introduced at a fortuitous time.

Barro's measure of the natural rate of unemployment increased by 2 percentage

and *FEDUN*, respectively. Technically, this is incorrect since the rational expectation of DM_t must be formed on an information set available in period t − 1. Using lagged values of these variables in the money equation weakens the power of both unemployment equations (4) and (8).

[9]In his equation (5), p. 108, Barro regressed the unemployment variable on the total rates of growth of money and the determinants of the natural unemployment rate. This provides a test of the effect of money in the *long run* since, according to the equation, a permanent shift from one money growth rate to another causes unemployment to change permanently.

[10]Between 1969 and 1970 Barro's fitted unemployment rate increased by 1.6 percent and the estimated natural unemployment rate by 2.0 percent. Since the fitted unemployment rate is the sum of the estimated natural unemployment rate and the effect of the monetary surprises, these figures indicate that the only reason the fitted rate increased was because the estimated natural rate rose; the effect of the monetary surprises must have been to hold the increase in the fitted rate below the increase in the estimated natural rate. Since the only other determinant of the natural rate, *MINW*, moved to decrease the natural rate over these years, the change in *MIL* is the reason for the increase in the estimated natural rate and thus also in the fitted unemployment rate. The years 1971, 1972, and 1973 tell basically the same story.

TABLE 1

Year	Job Leavers			Inductions	In School		
	Males (1)	Females (2)	M/F (3)	(4)	Males (5)	Females (6)	M/F (7)
1975	291	369	.789	0	3,927	3,803	1.033
1974	270	314	.860	0	3,601	3,586	1.00
1973	254	276	.920	36	3,762	3,582	1.050
1972	245	262	.935	27	3,827	3,674	1.042
1971	237	234	1.013	156	3,880	3,735	1.039
1970	209	214	.977	207	3,618	3,508	1.031
1969	164	171	.959	265	3,586	3,498	1.025
1968	167	167	1.000	340	3,503	3,504	1.000
1967	165	179	.922	298	–	–	–

Sources: Columns (1), (2), (5), and (6) are from the *Handbook of Labor Statistics 1976* columns (1) and (2) are from Table 56, p. 116, and columns (5) and (6) from Table 8, p 40 Column (4) is from *Statistical Abstracts of the United States 1976*, p. 339 All data is measured in thousands

points in 1970 and maintained this relatively high level thereafter. The effect of setting *MIL* equal to zero can conservatively be estimated as increasing the natural rate by 2 percentage points since the minimum wage, the other determinant of the natural rate, moved to decrease the natural rate over this period. Given a 1969 labor force of 84 million, this increase implies 1.68 million persons were unemployed at any given time throughout 1970 to 1973 because of the change in the military draft.

The rationale for setting *MIL* equal to zero after 1969 is that starting in 1970 either there were no draft calls or the calls were done under the lottery system. According to Barro, there were two effects on the labor market other than the direct conscription effect: a reduced incentive for students to remain in school and for employees to stay employed in order to avoid the draft. The magnitude of these effects implied by the change in Barro's measurement of the natural rate of unemployment is implausible.

The number of males and females who were job leavers and the number of inductees are listed in columns (1), (2), and (4), of Table 1. The effect of the change in the draft system on the job leavers should be reflected in a change in the relationship between the male and female components of Table 1. This should occur in 1970. The ratio of male to female job leavers should increase if a significant number of males did not leave their jobs prior to 1970 in order to avoid the draft. Little, if any, change in the ratio is apparent; see column (3) of Table 1.

Column (4) indicates that 605 thousand men were inducted in 1968 and 1969. Since inductions were for two years, the direct conscription effect of changing the draft in 1970 was equal to or less than this number.

In estimating the number of persons who enrolled in school to avoid the draft, two different data sets are available. One data set gives the number of persons who were not participating in the labor force because they were in school; see columns (5) and (6) of Table 1. The ratio of males to females in this data set should decrease in 1970 and afterwards if the draft induced male enrollments. The ratio actually increased, although insignificantly; see column (7) of Table 1.

Also, data are available on the percentage of persons of any age group who were enrolled in school. This data is given in Table 2. Comparison of the male and female series for either of the age groups shows that participation rates for males appears to have risen relative to females during the late 1960's. For the males 18-19 years of age, the increase to a 60.4 percent participation rate in 1969 is approximately 15 percentage points above the lowest rate which occurred in 1974. For the

TABLE 2

Year	18–19 Years of Age		20–24 Years of Age	
	Males	Females	Males	Females
1975	49.9	44.2	26.4	18 7
1974	45.8	40.7	25.8	17 3
1973	47.9	38.2	25.2	16 7
1972	51.2	41.8	27.8	16.0
1971	55.4	43 4	29.2	15 7
1970	54.4	41 6	29 3	15 2
1969	59.4	41.8	32 0	16.0
1968	60.4	41.3	30 5	14 3
1967	56.3	40.3	30 6	15 1
1966	57.8	37.7	29 2	12.4
1965	55.6	37 7	27 7	11 8
1964	50 9	33.7	23.8	10.9

Sources: Historical Statistics of the United States . . p. 370, and relevant issues of *Statistical Abstracts of the United States* for post-1970 data

males of 20–24 years, the maximum swing in the participation rate was 7 percentage points. Given a population of 3.3 million males aged 18–19 years and 3.9 million males of 20–24 years, these percentage increases imply 495 thousand males of 18–19 years and 266 thousand males of 20–24 years attended school to avoid the draft.

These calculations imply that, as a liberal estimate, 605 thousand inductees and 761 thousand students entered the ranks of the unemployed as soon as the draft ended. Clearly, not all these persons remained continually unemployed. The unemployment rate for Vietnam veterans was under 10 percent, for males aged 18–19 years under 15 percent, and for males of 20–24 years under 10 percent.[11] These unemployement rates indicate only 161,350 persons were added to the ranks of the unemployed. This is significantly less than Barro's estimate of 1.68 million persons added to the unemployment pool.

Barro has thus not given evidence that a large increase in the natural rate of unemployment is the reason for the rise in unemployment in the 1970's, and thus has no explanation for the later increase.

III. Conclusion

This note has examined Barro's attempt to empirically support the short-run rational expectations model of unemployment as formulated in equation (1). Barro's measurement of $(DM - DM^e)$, and thus also of deviations in unemployment from the natural rate, was examined in detail. In a formal test, Barro's monetary model was rejected in favor of a more flexible monetary model. After using two versions of the flexible model, and the two corresponding measurements of unanticipated monetary movements, the conclusion was reached that Barro's results do not present strong support for the rational expectations model of unemployment but rather indicate some support for the competing hypothesis that monetary policy is effective in the short run whether it is anticipated or not.

In Section II Barro's measurement of the natural unemployment rate was rejected. One of the determinants of the natural rate, the military draft variable, was found to have an implausible effect on the labor market. Thus the rational expectations hypothesis as represented by equation (1) has not been shown to be able to satisfactorily explain unemployment in the United States.

APPENDIX

Formally, the cost of a war during year i equals $X_i = MY \cdot (MIL_i - MIL_i^N)$, where MY = the cost per man-year = $C/[\Sigma(MIL_j - MIL_j^N)]$, the index j runs from the beginning to the end of the war, MIL_i = the number of military personnel on active duty during year i, MIL_i^N = the "normal" value of MIL_i, and C = the total cost of the war.

The MIL_i^N variable was derived by interpolating a constant rate of growth between normal prewar and postwar military personnel figures. For World War II and the Korean War these figures were the actual values for the years 1940, 1947, and 1950, 1955. For the

[11] *Sources: Handbook of Labor Statistics, 1967*, p. 114, for age group unemployment rates. For Vietnam veteran unemployment rates see various issues of the *Monthly Labor Review*, *Data*: Cost of World War II and the Korean War from *Historical Statistics . . .* , p. 1140. Original cost figures were used. Vietnam War cost from Robert Eisner's article, p. 118. Military personnel on active duty from *Historical Statistics . . .* , p. 1141

Vietnam War the 1965 military personnel figure was used through 1971 as the normal value since after the war the number of military personnel on active duty declined below the prewar value. Consecutive years were averaged in the Vietnam and Korean War years since the military figures are midyear data and the Korean War started and ended in a rather clear way at midyear.

Since a postwar figure is being used in the construction of MIL^N, information is given to the economic agents during wartime which they could not have possessed until after the war. Nonetheless, MIL^N was used because the use of the postwar data insured that the yearly cost of the war would smoothly approach zero by a time when it was clear that the war was over.

A second problem is the use of the total cost of war to arrive at the yearly cost of the war. Again, information is being used which was not available to the agents at the time. This is not a problem for the World War II series (*MILX*1) since *MY* does not depend on the yearly index i. Thus, if the total cost of World War II were left out of the X_is for World War II, the only affect would be on the regression coefficient of *MILX*1, and not on that coefficient's t-statistic or on any other part of the regression coefficient. With the Korean War and Vietnam War series, *MY* for the Korean War could be divided into the *MILX*2 series. Then the Korean War yearly data would not contain any information not available at the time, although this problem would still exist for the Vietnam War data since it would contain the ratio of *MY* for the Vietnam War to *MY* for the Korean War. This information could easily be derived at the beginning of the Vietnam War. Dividing total military expenditures for both 1966 and 1951 by the military personnel during those years and then taking the ratio of the former measure of the cost per man-year to the latter, a new measure of the ratio of the cost per man-year in the Vietnam War to the Korean War is obtained. This new figure is equal to 2.95, whereas the ratio of the *MY* terms is 3.13. Thus the information needed in measuring the impact of war was available.

Finally, the war series *MILX*1 and *MILX*2 are equal to the relevant X_i series divided by the money stock in that period.

The reason for the rather roundabout calculation using the deviation from normal of the military personnel rather than deviations of the military budget is twofold: the ideal measurement is of the cost of the war, and using military budget figures might introduce nonwar expenditures; secondly, the Vietnam War was not financed in the most public way possible. The military personnel figures are more easily and directly observable than the appropriate budget figures.

REFERENCES

R. J. Barro, "Unanticipated Money Growth and Unemployment in the United States," *Amer. Econ. Rev.*, Mar. 1977, *67*, 101–15.

R. Eisner, "The War and the Economy," in Sam Brown and Len Ackland, eds., *Why Are We Still in Vietnam?*, New York 1970.

R. E. Lucas, Jr., "Expectations and the Neutrality of Money," *J. Econ. Theory*, Apr. 1972, *4*, 103–24.

T. J. Sargent and N. Wallace, "'Rational' Expectations, the Optimal Monetary Instrument and the Optimal Money Supply Rule," *J. Polit. Econ.*, Apr. 1975, *83*, 241–54.

U.S. Bureau of the Census, *Historical Statistics of the United States, Colonial Times to 1970*, Washington 1975.

———, *Statistical Abstract of the United States, 1976*, Washington 1976.

U.S. Bureau of Labor Statistics, *Mon. Labor Rev.*, Washington various issues.

———, *Handbook of Labor Statistics 1976*, Washington 1976.

U.S. Department of Commerce, *Survey of Current Business*, Part 1, Washington, Jan. 1976.

Unanticipated Money Growth and Unemployment in the United States: Reply

By ROBERT J. BARRO*

David Small raises a number of questions concerning my 1977 study of money growth and unemployment in the United States. My reexamination of the evidence indicates that his major criticisms are unfounded—specifically, the estimated effects of monetary shocks on the unemployment rate are robust to his suggested changes in specification. On the other hand, there are some difficulties with the estimated unemployment effects of some "real" variables—that is, with the determinants of the "natural" unemployment rate.

I. Monetary Effects on Unemployment

The analysis in the first section of Small's comment amends my specification of money growth to include an additional influence during war years. The variable *FEDV*, which was intended to measure real federal expenditures relative to "normal," is allowed to have different coefficients for war and nonwar years. Even in Small's results, the incorporation of an extra effect of federal spending on money growth during wartime, as reflected in the coefficient on the *DFEDV* variable in his equation (3), does not alter the general nature of the impact of money shocks on unemployment in his equation (5). The changes that do occur, which are primarily a worsening of the fit of the unemployment equation, are a consequence of his puzzling treatment of World War II. The classification of "war years" over the 1941–73 sample for the money growth equation is (as indicated in his fn. 3) 1950–52, 1965–70. The United States was also involved in military disturbances from late 1941 to 1945. Small's procedure inexplicably treats the World War II observations in a group with the peacetime years; 1946–49, 1953–64, 1971–73; which are then separated from the Korean and Vietnamese War periods.[1]

I have redone the estimation of the money growth and unemployment rate equations with the wartime sample reclassified to 1941–45, 1950–52, and 1965–70. (Treatment of 1941 as a peacetime year has a negligible effect on the results.) Two other differences from Small's analysis are first, the sample endpoint is extended from 1973 to 1977 (which does not affect the conclusions), and second, there is an allowance for a higher variance of the error term for observations before 1946 in the money growth equation. As I discussed in my 1978 paper, weighting for this heteroscedasticity of the error term removes the apparent negative serial correlation of residuals in the estimated money growth equation, which eliminates the need for the Cochrane-Orcutt estimation procedure that Small applies. However, a negative verdict on Small's wartime modifications appears also in unweighted regressions.[2]

*University of Rochester. This research was supported by a grant from the National Science Foundation

[1]Small notes, "A dummy variable for World War II can be included with no significant effects on the results if the new dummy is entered independently of the dummy for the Korean and Vietnam wars, i.e., as long as the agents are not forced to believe all wars have the same impact on the money supply (p. 998, fn. 3)." First, the omission of a war dummy for the World War II years makes no sense, since the 1941–45 period is then treated as peacetime. Second, the use of a common dummy does not treat all wars equivalently, but rather measures the intensity of war by the federal spending relative to normal variable. Third, the logical extension of Small's multidummy methodology is to introduce a dummy for any observation for which the money growth residual has some explanatory power for unemployment—that would surely worsen the fit of the unemployment equation!

[2]One change in the money stock data should also be noted. The figures that I used earlier for M_1 before 1947 from Milton Friedman and Anna Schwartz (Table 2) do not include demand deposits due to foreign banks and some other adjustments that were made in revisions to the money stock concept in the October 1960 and August

With Small's wartime variable excluded, the estimated equation for money growth DM is

1941–77 sample (observations from 1941–45 weighted by 0.36):[3]

$$(1)\quad DM_t = \underset{(.024)}{.084} + \underset{(.14)}{.45}DM_{t-1} + \underset{(.12)}{.17}DM_{t-2} + \underset{(.016)}{.074}FEDV_t + \underset{(.008)}{.027} \cdot log(U/1-U)_{t-1}$$

$\hat{\sigma}$ (for post-World War II years) $= .0140$, $D.W. = 2.0$

where standard errors are shown in parentheses, $\hat{\sigma}$ is the standard error of estimate, and $D.W.$ is the Durbin-Watson statistic. (The Durbin h-statistic also shows no residual serial correlation for this equation.) The results for the unemployment rate are

1946–77 sample:

$$(2)\quad log\,U/(1-U)_t = \underset{(.13)}{-3.06} - \underset{(1.9)}{5.2}DMR_t - \underset{(1.9)}{11.8}DMR_{t-1} - \underset{(1.7)}{6.1}DMR_{t-2} - \underset{(0.7)}{4.0}MIL_t + \underset{(0.5)}{1.3}MINW_t$$

$$R^2 = .80,\ \hat{\sigma} = .14,\ D.W. = 2.0$$

where DMR_t—a measure of "unanticipated money growth"—is the contemporaneous residual from equation (1), MIL is a military personnel variable, and $MINW$ is a minimum wage rate variable.[4]

With the wartime variable $DFEDV$ included, the estimated money growth equation becomes

1941–77 sample (observations from 1941–45 weighted by 0.36):

$$(3)\quad DM_t = \underset{(.022)}{.095} + \underset{(.14)}{.43}DM_{t-1} + \underset{(.12)}{.13}DM_{t-2} + \underset{(.023)}{.047}FEDV_t + \underset{(.025)}{.039}DFEDV_t + \underset{(.008)}{.029} \cdot log(U/1-U)_{t-1}$$

$$\hat{\sigma} = .0137,\ D.W. = 2.1$$

where $DFEDV$ is the $FEDV$ variable multiplied by a dummy that takes on the value one for the "war years," (1941–45, 1950–52, 1965–70), and zero otherwise. The unemployment rate results are then

1946–77 sample:

$$(4)\quad log\,U/(1-U)_t = \underset{(.16)}{-3.05} - \underset{(2.5)}{5.5}DMR_t^* - \underset{(2.4)}{11.7}DMR_{t-1}^* - \underset{(2.1)}{7.5}DMR_{t-2}^* - \underset{(0.9)}{4.7}MIL_t + \underset{(0.6)}{1.5}MINW_t$$

$$R^2 = .72,\ \hat{\sigma} = .16,\ D.W. = 1.7$$

where DMR_t^* is the contemporaneous residual from equation (3).

The estimated coefficient of Small's $DFEDV$ variable in equation (3) has a t-value of 1.6, which does not indicate significant difference from zero according to usual statistical standards. The effect of including the $DFEDV$ variable in the money growth equation on the estimates for the unemployment rate are minor. There is some worsening of the fit in equation (4) relative to that in equation (2), but there are no important changes in the estimated coefficients or their standard errors.

The results raise a general issue concerning the interaction between the money growth

1962 issues of the *Federal Reserve Bulletin*. Because these revisions were applied retroactively to data since 1947, but not before that date, the effect is a one-time jump of about 1.5 pecent in the level of money in 1947. The data on the money stock for the present analysis contain a corresponding upward adjustment to the level of money prior to 1947 in order to eliminate this problem. This modification of the data is not important for the principal issues discussed in this paper—see, however, fn 10 below for one effect of the change.

[3]The weight 0.36 is determined from a maximum-likelihood criterion.

[4]The $MINW$ variable is defined as the ratio of the applicable federal minimum wage rate to average nonfarm gross hourly earnings, multiplied by the ratio of workers covered by federal minimum wage standards to total employment. The measure has been revised from that in my 1977 paper (Table 2) to allow for two categories of minimum wage rates and to use total employment rather than nonsupervisory employment in the denominator. However, the old and new series for $MINW$ are closely correlated and this change does not materially affect any results.

and unemployment equations. It is conceivable that a variable, such as *DFEDV*, would have a marginally significant effect when included in the money growth equation, but would nevertheless have a substantial effect on the pattern of monetary residuals and, accordingly, a strong—possibly adverse—influence on the estimated equation for the unemployment rate (or output or the price level). The present two-part estimation procedure gives no weight to this latter effect in obtaining estimates of the coefficients of variables in the money growth equation. An efficient procedure would estimate the money growth and unemployment (and other) equations jointly.[5]

Non-linear joint maximum-likelihood estimates (from the Time-Series Processor regression package) of the money growth and unemployment equations turn out to be, with the *DFEDV* variable excluded,

1941–77 sample (observations from 1941–45 weighted by 0.36):

$$(1') \quad DM_t = \underset{(.012)}{.075} + \underset{(.11)}{.36DM_{t-1}} + \underset{(.09)}{.15DM_{t-2}} + \underset{(.011)}{.088FEDV_t} + \underset{(.004)}{0.23} \cdot log(U/1-U)_{t-1}$$

$$\hat{\sigma} = .0134, \; D.W. = 1.8$$

1946–77 sample:

$$(2') \quad log\, U/(1-U)_t = -\underset{(.15)}{2.84} - \underset{(1.5)}{5.1DMR_t} - \underset{(1.5)}{116.DMR_{t-1}}$$

$$\underset{(1.8)}{6.5DMR_{t-2}} - \underset{(0.8)}{5.6MIL_t} + \underset{(0.5)}{0.7MINW_t}$$

$$\hat{\sigma} = .10, \; D.W. = 2.5$$

where DMR_t is the contemporaneous residual from equation (1′) and asymptotic standard errors are shown in parentheses. Note that the reported standard errors of estimate are not adjusted for degrees of freedom as in equations (1) and (2). The main changes from the two-part estimates shown in equations (1) and (2) are: first, as would be expected, a small worsening in fit of the money growth equation (recall that $\hat{\sigma}$ in equation (1′) is not adjusted for degrees of freedom) and improvement of fit of the unemployment equation; second, a reduced coefficient of the minimum wage rate variable *MINW* in the unemployment equation (see below); third, a substantial reduction in the estimated standard error of the lagged unemployment rate variable in the money growth equation (but little change in the point estimate); and fourth, a reduced coefficient of DM_{t-1} in this equation.

If the *DFEDV* variable is included in the money growth equation, its estimated coefficient from the joint maximum-likelihood procedure is .002, standard error = .011, and the remainder of the results are virtually unchanged from those shown above. Therefore, in the joint estimation that considers the indirect effect of money growth estimates on the calculation of the *DMR* variable and the consequent effect on the fit of the unemployment rate equation, the *DFEDV* variable has a negligible overall influence.[6]

The analysis at the end of Section I of Small's comment makes little sense to me. First, the two military variables in his equa-

[5]See Leonardo Leiderman and my paper with Mark Rush on this matter.

[6]This joint estimation is subject to the cross-equation restrictions that monetary effects enter into the determination of unemployment only as the residual, $DMR = DM - \hat{DM}$, where $\hat{DM}$ is the estimated value from the money growth equation. This hypothesis can, of course, be tested within the framework of joint maximum-likelihood estimation. With the *DFEDV* variable excluded from the *DM* equation, the appropriate likelihood ratio test yields a statistic (for $-2 \cdot log$ (likelihood)) of 13.2, which is below the 5 percent critical value for the χ^2 distribution with eight degrees of freedom of 15.5. With the *DFEDV* variable included, the statistic is 18.4 with a 5 percent χ^2 value with eleven degrees of freedom of 19.7. Therefore, the inclusion of the *DFEDV* variable does not affect the outcome of this statistical test. With or without the inclusion of this variable, the likelihood ratio test statistic turns out to be much smaller if an output equation (over 1946 to 1977) is considered rather than an unemployment rate equation, or if the sample for the unemployment rate begins in 1949 rather than 1946. There do appear to be some problems with the estimated unemployment rate equation for the immediate post-World War II years (see the discussion below).

tion (7) seem principally to be designed as *ex post* mechanisms for getting the wartime monetary residuals close to zero. Second, it would seem interesting to construct a model of federal expenditure in order to divide this variable into anticipated/unanticipated and temporary/permanent components and thereby to improve on my *FEDV* variable. (From previous examination of federal spending, I concluded that this exercise would amount mostly to forming predictions of the occurrence, size, and duration of wars.) Small's procedure identifies "unanticipated" federal spending with the residual from an autoregression, his equation (6), that includes also a lagged value of *GNP*. The analysis is flawed by the measurement of federal spending in equation (6) in nominal terms—except for brief periods it would seem more reasonable to model the determination of this spending in real terms.[7] Consider the effect of an *unanticipated* increase in the price level, which would be associated with an unanticipated increase in the money stock or other forces. If real federal spending over the year is unchanged, the price increase would lead to a corresponding rise in nominal federal expenditure, which Small would measure as a higher value of unanticipated federal spending (*FEDUN*). According to his equation (7), there would be a corresponding rise in the measured value of "anticipated" money growth—in other words Small's procedure filters out a substantial portion of unanticipated price and money change from his concept of "unanticipated" money growth *DMT*. Given this measurement problem and the *ex post* nature of the two military variables in his equation (7), it is hard to see what theory is being tested in his equation (8).

II. Determinants of the Natural Unemployment Rate

Small expresses some doubts concerning my estimates of the effects of "real" variables on unemployment. A difficulty that he does not raise concerns the estimated minimum wage rate effect, which does not stand up to further investigation. The joint estimates shown above in equation (2′) and results from quarterly data (to be reported) indicate that the positive unemployment effect of the *MINW* variable, as shown in equations (2) and (4), is unreliable. In particular, the effect of the *MINW* variable is insignificant if the sample for the unemployment equation is begun in 1949, rather than 1946. Measuring *DMR* values as the residuals from equation (1), the unemployment rate equation over 1949 to 1977 is

1949–77 sample:

$$(5)\quad \log U/(1-U)_t = -\underset{(.04)}{2.68} - \underset{(1.6)}{4.6}DMR_t - \underset{(1.6)}{11.0}DMR_{t-1} - \underset{(1.6)}{5.7}DMR_{t-2} - \underset{(0.6)}{5.4}MIL_t$$

$$R^2 = .87,\ \hat{\sigma} = .11,\ D.W. = 2.5$$

If the *MINW* variable is added to equation (5), its estimated coefficient is −0.1, standard error = 0.6.[8] A statistical test that the 1946–48 observations are generated from the same model (with the *MINW* variable included) as that applicable to the 1949–77 sample yields the statistic,[9] $F^3_{24} = 5.3$, which exceeds the 5 percent critical value of 3.0. The suggestion is that the estimated positive influence of the *MINW* variable on unemployment was merely an imperfect attempt to account for the otherwise unexplained low values of the unemployment rate from 1946 to 1948.[10] While this problem may involve some persisting effects of World War II, it is noteworthy that the *MINW* variable is insig-

[7]Small says "Equation (6) can be run in real terms with no significant effect on the results" (p. 999, fn. 6). This "robustness" during a period of rising inflation suggests that the federal spending equation is in fact not well specified.

[8]A similar result appears in jointly estimated money growth and unemployment equations, with the unemployment sample limited to 1949–77

[9]A difficulty with this test is that the hypothesis was not generated independently of the data.

[10]From equation (1), the actual and estimated values for the unemployment rate are for 1946, .042 vs. .043; for 1947, .038 vs. .045; and for 1948, .037 vs. .051. The discrepancies for 1947 and 1948 were increased by the—apparently warranted—adjustment in the money stock series prior to 1947 (see fn. 3 above).

nificant in an equation for real *GNP* even for samples that begin in 1946. Further, the 1946–48 observations on *GNP* can satisfactorily be grouped with the 1949–77 values.

Small reasonably questions my military personnel/conscription variable *MIL*. In my 1978 paper I noted some doubts about this variable, which concerned its surprisingly strong output effect and insignificant price level influence. Although the *MIL* variable is highly significant in unemployment equations, even in samples that begin in 1949 (see equation (5) above), it should be noted that this variable, as shown in Table 2 of my 1977 paper, does not exhibit major variations from 1951 to 1969, especially from 1955 to 1969. Mostly, the *MIL* variable shows a sharp increase from its 1949–50 values at the start of the Korean War, a mild decline from 1953 to 1958, a mild increase with the Vietnam War for 1967–69, and a sharp drop (to zero with the end of the selective, nonlottery draft) in 1970.

The possibility that the military personnel variable is proxying for movements in federal purchases can be considered in an equation that substitutes the ratio of real federal purchases of goods and services to real *GNP*, G/y, for the *MIL* variable. Again using the residuals from equation (1) to measure the *DMR* values, the estimated equation is[11]

1949–77 sample:

$$(6)\ \log U/(1-U)_t = \underset{(.12)}{-2.20} - \underset{(2.0)}{6.3}DMR_t - \underset{(2.0)}{10.5}DMR_{t-1} - \underset{(2.0)}{2.0}DMR_{t-2}) - \underset{(1.0)}{6.8}(G/y)_t \qquad \hat{\sigma} = .14,\ D.W. = 1.7$$

Lagged values of G/y are insignificant when added to equation (6). The estimated equation does suggest an important expansionary effect of the contemporaneous amount of federal purchases. (Another result from equation (6) is the loss of significance of the DMR_{t-2} variable—that is, with G/y substituted for *MIL*, the lagged effect from money shocks to unemployment is shorter than that estimated previously.) However, if the *MIL* variable is added to equation (6), its estimated coefficient is significant (−4.5, standard error = 1.4), while that on G/y becomes insignificant (−1.4, standard error = 1.8). Similar results obtain even if the sample is limited to 1949–69; that is, if the period where the *MIL* variable drops to zero is omitted. It turns out that the superior performance of the *MIL* variable arises also if real *GNP* is used instead of the unemployment rate as a dependent variable. At this point my analysis lacks a fully satisfactory specification of the military/government purchases effect on the unemployment rate and output.

It may be worth noting that equation (5), which includes the *MIL* variable, and equation (6), which contains G/y, have similar implications for the time path of the natural unemployment rate. With all *DMR* values and the error term set to zero, equation (5) implies an unemployment rate of 6.4 percent at the 1977 value of *MIL* (zero), and a value of about 4.5 percent for the values of *MIL* (.07 to .08) prevailing in the early 1960's. Equation (6) yields values for the unemployment rate of 6.2 percent at the 1977 value of G/y (.076) and also about 4.5 percent for the values of G/y (around .125) that existed in the early 1960's. Conceivably, this pattern for the natural unemployment rate is approximately correct even if neither the *MIL* nor the G/y variable are the properly specified military/government purchases influence on unemployment.

III. Conclusions

My principal conclusion is that the estimated effects of monetary disturbances on unemployment are robust to Small's changes in the specification of the money growth

[11] The estimation uses $G/\hat{y}$ as an instrument for G/y, where $\hat{y}$ is an estimated value of real *GNP* from a separate equation. Because short-run movements in G are dominated by shifts in defense spending, I have not considered the possibility of reverse feedback from economic activity to G. Since government transfers are not included in G, this feedback is probably unimportant.

ocess. Accordingly, the principal business cle results from my earlier studies remain tact. The analysis of "real" determinants of nemployment is less secure. The previously stimated minimum wage rate effect is unreble. The estimated military personnel/fedal purchases influences have also not been rmly established.

REFERENCES

. J. **Barro,** "Unanticipated Money Growth and Unemployment in the United States," *Amer. Econ. Rev.*, Mar. 1977, *67*, 101–15.

——, "Unanticipated Money, Output, and the Price Level in the United States," *J. Polit. Econ.* Aug. 1978, *86*, 549–80.

—— **and M. Rush,** "Unanticipated Money and Economic Activity: Results from Annual and Quarterly U.S. Data," paper presented at Nat. Bur. Econ. Res. Conference on Rational Expectations and Economic Policy, New Hampshire, Oct. 1978.

Milton Friedman and Anna J. Schwartz, *Monetary Statistics of the United States*, New York 1970.

L. Leiderman, "Macroeconometric Testing of the Rational Expectations and Structural Neutrality Hypotheses for the United States," *J. Monet. Econ.*, Jan. 1980, *6*.

David H. Small, "Unanticipated Money Growth and Unemployment in the United States: Comment," *Amer. Econ. Rev.*, Dec. 1979, *69*, 996–1003.

Professorial Behavior Given a Stochastic Reward Structure

By WILLIAM E. BECKER, JR.*

Concern over faculty collective bargaining, as opposed to salary assignment on individual merits, has resulted in studies on the effects of unionization on faculty behavior. Pressure for the application of equal pay regulation, instead of comparison of individuals, has also led to studies on the faculty behavior effects of affirmative action. Similarly, desire for a community of scholars, versus a system of competition between scholars, has fostered studies on faculty response to teaching and research evaluation. As Howard Tuckman (p. 40) notes, however, there has been little serious effort in any of these studies to construct mathematical models of faculty behavior along the optimization lines favored by many economists. Without such modeling one can only speculate on the faculty productivity consequences of changing methods for determining faculty rewards.

In this study a constrained utility optimization model is presented as the behavior model of a university professor. Following my earlier work, it is assumed that faculty time is a variable input to various academic outputs. Academic outputs generate both personal satisfaction and income. As in Joseph Stiglitz's theory of screening and income assignment, it is assumed that academic outputs may not be precisely measurable by the university.[1] Where there is little screening of faculty productivity, all faculty within a given error of measurement receive the same salary regardless of their actual contribution. If screening is accurate, however, individuals are shown to receive the value of their output.

Output effects resulting from changes in the intensity of screening are shown to depend on the faculty member's position in the output distribution. For high producers, changes in measurement accuracy will result in like changes in the desire to produce. Furthermore, changes in pecuniary return to a given academic output will produce like changes in the professor's desire to produce that output only if screening methods are highly accurate.

I. Professorial Behavior and Institutional Arrangements

It is assumed that a professor is an individual who not only derives utility from consumption activity, but also acquires satisfaction from the professional activities of teaching and research. The professorial utility function is therefore written as

$$(1)\qquad U(Q_1, Q_2, Q_3),\ U_i = \partial U/\partial Q_i \geq 0$$

where Q_i, $i = 1, 2, 3$ are, respectively, teaching output, research output, and consumption.[2] The professor is assumed to know the exact quality and quantity of his or her professional output. Consumption, on the other hand, is partially dependent on the university's measurement of the professor's

*Indiana University. Discussions with Frank Boddy, Jim Henderson, Herb Mohring, and members of the Spring 1977 Minnesota Applied Microeconomic Workshop and Fall 1977 Minnesota Academic Rewards Structure Symposium are gratefully acknowledged. The managing editor of this *Review* and an anonymous referee provided helpful clarifying suggestions. Jack Rodgers and Dick Davis provided constructive criticism in rechecking my algebraic manipulations and implications.

[1]University is used here as a generic term representing a college, department, or other nonprofit-oriented academic unit which may make salary decisions. It is also used, however, to exclude certain institutions such as junior colleges which give no emphasis to research credentials.

[2]Douglas Needham and I in separate studies provide constructive criticism of the elementary, two-dimensional indifference curve analysis of faculty behavior as presented by Richard McKenzie. The restriction to only two professional outputs and a single consumption output is not essential to this analysis. Different forms of faculty output can be separated out, and one can consider the disaggregation of time spent in each activity as the appropriate time input.

professional output.[3]

The professor must produce professional output through the input of time. The amount of time spent in teaching activity affects production of Q_1; time devoted to research activity affects the production of Q_2. However, time spent in research activity may enhance teaching quality since it allows the professor to share disciplinary experiences with students. The time spent in discipline-based teaching may also sharpen the quality of the professor's research.[4] The professor's production process, represented for simplicity as linear in time, is therefore given by

$$(2) \qquad Q_i = \sum_{j=1}^{2} a_{ij} T_j, i = 1, 2$$

where the T_j, $j = 1, 2$ are, respectively, time spent in teaching and research activity. The technology coefficients for a fixed capital stock (both human and physical) are assumed to be $a_{ii} > a_{ij} \geq 0$, $i \neq j$.

For consumption, the professor needs income to purchase market goods. He also needs leisure-consumption time to acquire and enjoy goods purchased. The consumption-production function is therefore written linearly as

$$(3) \qquad Q_3 = bY + a_{33} T_3$$

where Y is university-determined income and T_3 is time spent in leisure-consumption activity. Coefficients b and a_{33} are assumed to be positive.

It is assumed that the professor obtains personal income from the university on the basis of teaching and research output. Likewise, the university generates its revenue only on the basis of total institutional outputs of teaching and research. The link between the university generation and faculty receipt of income need not be direct, however.

Universities may affirm an intellectual value structure that presupposes little inherent superiority of knowledge in various fields.[5] Differences in the quality and quantity of faculty performance may be ignored in the university's judicating. Superficial judging of faculty output results in individuals of similar experience, academic rank, or task assignment being labeled equally productive even though their actual contribution to the university's output may differ.[6] Such incorrect labeling implies that, in essence, the high performers in a given group of scholars are sharing their productivity with the less able.

The extent to which individuals share productivity because of incorrect labeling is a function of the university's screening process applied to teaching and research output. To see this, consider an example in which the number of students enrolled in undergraduate honors thesis seminars is believed to net the university, after all nonfaculty expenses, W dollars per student thesis completed.

In the extreme case where the university employs no screening of faculty thesis contribution, all professors must be labeled as equally productive. With the exception of the truly average professors, all faculty are incorrectly labeled. To exhaust the university product, all faculty would receive the mean of total faculty product, regardless of actual contribution to student theses.[7] The high producers in the group subsidize the low producers.

If, after a somewhat intensive review of recorded thesis advising, the university

[3] It is assumed that the professor is risk neutral in regard to measurement error.

[4] For example, in the foreword to Michael Evans' macro theory text, Lawrence Klein cites the importance of creative research as an input to quality teaching. Martin Bronfenbrenner, in the preface of his authoritative book on income distribution, acknowledges the importance of teaching as an input to research.

[5] Richard Freeman (p. 105) found that the interfield coefficients of variation are far lower in academic than in industry or government pay structures. On the other hand, many studies reviewed by Tuckman do show a statistically significant contribution for different disciplines to faculty salary after controlling for teaching, research, and other specific faculty and university output measures. However, the fixed difference in salary they attribute to the discipline of the professor typically amounts to less than $3,000 per year.

[6] For example, student credit hours generated by a faculty member are seldom used in determining the faculty member's salary while they do generate, via some formula, revenue for the department or university as a whole. The individual contributions of scholars of jointly authored articles or the contribution of those recognized by an author for providing constructive suggestions in preparing an article are seldom recognized.

[7] The university, by definition, is a nonprofit entity which cannot have a surplus or deficit.

decides there are three identifiable ranges of thesis contribution, then a low, medium, and high labeling is possible. The university can now pay each professor the mean product value for the group to which he or she is assigned. However, all professors receiving one of these labels need not be equally productive. Some may be at the high end of the productivity range within each of the three separate groupings and some may be at the low end. Those at the high and low ends of each group range are mislabeled and incorrectly paid, but not as badly as when there was no screening. Through yet more intensive screening, the university could continue to reduce the measurement error and, at least conceptually, get to the point of perfect labeling of all professors.

In general, a university's screening process of intensity k_1 for teaching and k_2 for research output can be thought of as a point estimate of a professor's productivity. Let $e_i(Q_i, Q_i^*, k_i)$ be the probability that a professor of productivity Q_i is labeled Q_i^* in a university screening of intensity k_i. As k_i is increased the probability of correct labeling in the ith output also rises; i.e.,

$$(4) \quad \frac{\partial e_i(Q_i, Q_i^*, k_i)}{\partial k_i}\bigg|_{Q_i^* = Q_i} \geq 0, \; i = 1, 2$$

Let $h(Q_i)$ be the unimodal density function of Q_i, that is, percentage of faculty whose productivity is Q_i. As in the example, those faculty labeled Q_i^* in a screening of intensity k_i receive an income for this output equal to their mean product value. The algebraic expression for this payment mechanism is now given by

$$(5) \quad Y_i(W_i, k_i, Q_i^*) = W_i \int Q_i e_i(Q_i, Q_i^*, k_i) h(Q_i)\, dQ_i \div \int e_i(Q_i, Q_i^*, k_i) h(Q_i)\, dQ_i$$

where W_i is the dollar weight given to the ith output.

Multiplying both the numerator and the denominator of (5) by the number of professors in the university would show the numerator to be the total product value of all those labeled Q_i^* and the denominator to be the number of faculty so labeled. A professor whose true productivity is Q_i will have an expected income from this output given by

$$(6) \quad Y_i(W_i, k_i, Q_i) = \int Y_i(W_i, k_i, Q_i^*)\, e_i(Q_i, Q_i^*, k_i)\, dQ_i^*$$

The income consequence of any imprecise, yet somewhat accurate, assessment of a professor's position in the distribution of outputs can be established by considering a screening process in which

$$(7) \quad e_i(Q_i, Q_i^*, k_i) = f_i(Q_i - Q_i^*, k_i) = f_i(\epsilon_i, k_i)$$

$$(8) \quad \begin{aligned} E(\epsilon_i) &= 0 && \\ E(Q_i \epsilon_j) &= 0, && i = j, i \neq j \\ E(\epsilon_i^2) &= g_i(k_i) > 0, && g_i' < 0 \\ E(\epsilon_i \epsilon_j) &= 0, && i \neq j \end{aligned}$$

The assumption of a fairly accurate screening system as given in (7) and (8) implies that (5) and (6) can be approximated, summing over both outputs, as[8]

$$(9) \quad Y(W, k, Q^*) \simeq \sum_{i=1}^{2} W_i [Q_i^* + h_i'(Q_i^*) g_i(k_i)/h_i(Q_i^*)]$$

$$(10) \quad Y(W, k, Q) \simeq \sum_{i=1}^{2} W_i [Q_i + h_i'(Q_i) g_i(k_i)/h_i(Q_i)]$$

Equation (10) demonstrates that if the university screening of faculty is perfect, $g_i(k_i) = 0$ and $Q_i^* = Q_i$, a professor can expect to receive his or her own product value $W_i Q_i$. If screening is not perfect, $g_i(k_i) > 0$, a professor below the mode $h'(Q_i) > 0$ can expect to receive more than under perfect screening; a professor above the mode $h'(Q_i) < 0$ can expect to receive less.[9] This is a consequence of faculty being grouped within a given range of error with some colleagues who are better, but underrated, and some who are worse, but overrated. Below the modal output faculty tend to be grouped with more

[8] Using a Taylor expansion it can be shown that $h_i(Q_i + \epsilon_i) \simeq h_i(Q_i) + h_i'(Q_i)\,\epsilon_i$.

[9] An error in measurement problem, $g_i \neq 0$, is typically acknowledged but then ignored in empirical estimation of faculty salary-academic output regressions of the type given in equations (9) and (10).

who are better than worse; above the modal output the reverse is true.

The professor's decision-making framework can now be described by the algebraic expressions given above. The professor attempts to maximize utility, (1), subject to the production functions (2) and (3), expected income (10) and the total time available T in the period under consideration, that is,

$$(11) \qquad T = \sum_{i=1}^{3} T_i$$

In this decision framework, the professor is free to invest time in research and teaching. Increase in activities, via the production functions represented in (2), raises the professor's level of satisfaction through the utility function (1). The change in income resulting from increased output depends on the professor's position in the given output distribution and on the university's accuracy in measuring productivity as given in (10). If an increase in output results in an increase in expected income, then the professor's level of satisfaction via increased consumption in (3) is possible.

The time constraint, however, limits the level of satisfaction obtainable. An increase in the time invested in professional activity, while raising academic output and expected income, forces a reduction in the time spent in consumption activity. A possible resulting reduction in consumption lowers the professor's level of satisfaction.

The equilibrium of teaching, research, and consumption outputs depends upon the marginal rate of substitution in the utility function and production functions, and on the income determination function. It also depends on the distribution of the two professional outputs, the faculty member's position in those distributions and the variance connected with output measurement utilized by the university.

II. Equilibrium Output of Teaching, Research, and Consumption

To solve for the equilibrium level of Q_i, the decision framework is simplified to an objective function involving only one constraint. Since the utility function is defined in output space while the time constraint is not, the time constraint must be mapped into output space. This is possible since expected income can be expressed in terms of time spent in professional activities. Substituting (2) into (10) transforms the income determination function into time space.

$$(12) \qquad Y(W, k, Q) = \sum_{j=1}^{2} \Big[\Big(\sum_{i=1}^{2} W_i a_{ij} T_j\Big) + W_j h'_j g_j / h_j\Big]$$

Substituting (12) into (3) eliminates income from the equation system and permits expression of the three production functions in matrix form where Q_1, Q_2, and Q_3 can be seen as direct functions of T_1, T_2, and T_3 for a given distribution of faculty output and variance in measuring output:

$$(13) \qquad q = \begin{bmatrix} Q_1 \\ Q_2 \\ \bar{Q}_3^* \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ a_{31} & a_{33} & a_{33} \end{bmatrix} \cdot \begin{bmatrix} T_1 \\ T_2 \\ T_3 \end{bmatrix} = A \cdot t$$

where $a_{3j} = b \sum_{i=1}^{2} W_i a_{ij}$ $\qquad j = 1, 2$

$$\bar{Q}_3 = Q_3 - b \sum_{j=1}^{2} W_j h'_j g_j / h_j$$

Multiplying both sides of production system (13) by the inverse of A and then substituting into (11) gives the time constraint in output space:

$$(14) \qquad T = \sum_{i=1}^{3} \pi_i Q_i$$

where
$$\pi_1 = \frac{a_{22} - a_{21}}{a_{22}a_{11} - a_{12}a_{21}} - \frac{bW_1}{a_{33}}$$

$$\pi_2 = \frac{a_{11} - a_{12}}{a_{22}a_{11} - a_{12}a_{21}} - \frac{bW_2}{a_{33}}$$

$$\pi_3 = \frac{1}{a_{33}}\Big(1 - \frac{b}{Q_3} \sum_{i=1}^{2} W_i h'_i g_i / h_i\Big)$$

All three πs are expressed in terms of the

time per unit of output. The terms π_1 and π_2 are shadow prices of respective academic output while π_3 can be thought of as a "quasi-shadow price" since its value depends on values of the Q_is. A sufficient condition for $\pi_3 > 0$ is that $h_i' < 0$; a condition met if the professor is above the mode output for Q_i, $i = 1, 2$. A necessary condition for $\pi_i > 0$, $i = 1, 2$, is that the technology matrix represented by (2) is positive definite with dominant diagonal elements. Sufficient conditions to insure that (14) is negatively sloped and concave to the origin are

$$\pi_i > 0 \qquad i = 1, 2, 3 \tag{15}$$

$$h_i''h_i - h_i'^2 < 0 \qquad i = 1, 2 \tag{16}$$

$$h_i^2 h_i''' - 3h_i h_i' h_i'' + 2h_i'^3 \leq 0 \qquad i = 1, 2 \tag{17}$$

To avoid the possibility of corner equilibrium solutions, situations not normally observed in the real world, conditions (15) through (17) are assumed to hold. Conditions (16) and (17) can hold for a bell-shaped output density function which has one inflection point below the modal output and one inflection point above the modal output. Such a density function need not be symmetric.

The professor's decision-making framework is now summarized by

$$U(Q_1, Q_2, Q_3) + \lambda\left(T - \sum_{i=1}^{3} \pi_i Q_i\right) \tag{18}$$

Differentiating (18) with respect to Q_i, $i = 1, 2, 3$, gives

$$U_i - \lambda\left[\pi_i - \frac{h_i''h_i - h_i'^2}{a_{33}h_i^2} g_i bW_i\right] = 0 \qquad i = 1, 2 \tag{19}$$

$$U_3 - \lambda \frac{1}{a_{33}} = 0$$

$$T - \sum_{i=1}^{3} \pi_i Q_i = 0$$

Equation system (19) defines the equilibrium level of teaching, research and consumption desired and producible by the professor.

III. Reward Structure and Screening Intensity Changes

To establish the effect of a change in the income weight given to an output versus a change in the university's effort to measure that output, it is necessary to differentiate the first-order equation system (19) with respect to W_i and k_i. This differentiation yields

$$Hd = w \tag{20}$$

where H is a symmetric bordered Hessian matrix whose off-diagonal element in the ith row and jth column $(i, j = 1, 2, 3)$ is the second-order partial U_{ij}. The first two diagonal elements are $U_{ii} + g_i\lambda bW_i(h_i^2h_i''' - 3h_ih_i'h_i'' + 2h_i'^3)/a_{33}h_i^3$; the third diagonal element is U_{33}. The fourth column and row has a zero corner element and elements $1/a_{33}$ and $\pi_i - (h_i''h_i - h_i'^2)\,bg_i W_i/a_{33} h_i^2$, $i = 1, 2$. The matrix H is written to conform to the column vector d, written in horizontal position as

$$d = [\partial Q_1, \partial Q_2, \partial Q_3, -\partial\lambda]$$

The column vector w is made up of the following elements in descending order.[10]

$$w_i = \frac{-\lambda b(h_i''h_i - h_i'^2)}{a_{33}h_i^2} W_i g_i' \partial k_i - \frac{\lambda b}{a_{33}} \cdot \left(1 + \frac{h_i''h_i - h_i'^2}{h_i^2} g_i\right) \partial W_i \qquad i = 1, 2$$

$$w_3 = 0$$

$$w_4 = -\sum_{i=1}^{2} \frac{h_i' b}{a_{33}h_i^2} W_i g_i' \partial k_i - \frac{b}{a_{33}} \cdot \sum_{i=1}^{2} \left(Q_i + \frac{h_i'}{h_i} g_i\right) \partial W_i - \partial T$$

In vector w, the marginal utility of total time λ is positive if the marginal utility of the three utility arguments are positive as assumed. Condition (16) and $g_i' < 0$ gives sign determinancy of the multiplicative factors of ∂k_i. This sign determinancy, however,

[10] In the differentiation of equation (20) with respect to k_i, it is assumed that the change in screening intensity does not require a change in the nonproductive time of the professor being evaluated. For example, it is assumed that the professor does not have to invest his or her own time to administer and score these survey forms. If there is any pecuniary fee connected with output screening, an article submission fee for example, then W_i is interpreted as being net of this fee. Changes in faculty "signalling" are assumed to be unreadable by the university unless accompanied by changes in the intensity of university screening.

depends on whether the professor is above or below the faculty modal output. If he or she is above the mode, then the multiplicative factors of ∂k_t in w_1 and w_4 are negative. If he or she is below the mode, then w_4 shows a multiplicative factor for ∂k_1 which is positive; however, the multiplicative factor of ∂k_1 in w_1 continues to be negative. Sign determinancy of the multiplicative factors of ∂W_t in vector w is not possible unless g_i is small.

Kenneth Eble (p. 58) speculated that improving the system of evaluating teaching would lead to improved instruction. Similarly, but in a reverse direction, Richard Lester (pp. 130–31) speculated that the application of equal pay regulation would destroy university methods of screening faculty and thus threaten the productivity of faculty. The feasibility of their speculations can be explored by multiplying both sides of (20) by the inverse of H. Denoting the determinant of H by $|H|$ and the cofactor of the element in the ith row and jth column of H by H_{ij} $(i, j, = 1, 2, 3, 4)$, then multiplying both sides of (20) by the inverse of H and solving for $\partial Q_t/\partial k_t$ yields the change in teaching output resulting from a change in screening.

$$\frac{\partial Q_1}{\partial k_1} = \frac{-\lambda H_{11} b(h_1'' h_1 - h_1'^2) W_1 g_1'}{|H| a_{33} h_1^2} - \frac{H_{41} b h_1' W_1 g_1'}{|H| a_{33} h_1^2} \tag{21}$$

Given the specified set of sufficient conditions for a constrained utility maximum so that $|H| < 0$ and $h_{11} > 0$, the sign of (21) can be made signwise determinant by assuming normality that is $\partial Q_1/\partial T = (-H_{41}/|H|) > 0$. If the professor is above the modal teaching output, then $\partial Q_1/\partial k_1 > 0$. If the professor is below the mode, however, the substitution effect (first term in (21)) is of a different sign than the output effect (second term in (21)). The sign $\partial Q_t/\partial k_t$ is, therefore, indeterminate. (Similar results follow for the other academic outputs.)[11]

The sign determinance of (21), for those above the modal output, and indeterminance for those below, can be seen by noting that in equation (14) an increase in the screening of teaching results in a decrease in the "price" of consumption output for those above the teaching mode, and an increase for those below. At any given consumption output, for those above the teaching mode, the teaching-research output frontier shifts up, while for those below, this frontier shifts down. But the shift in this frontier is not parallel since, from the first-order conditions in (19), the production of teaching becomes relatively less expensive, regardless of the professor's position on the output distribution. Therefore, for those above the teaching mode, the output and substitution effects reinforce each other resulting in an increase in teaching output. On the other hand, for those below the teaching mode, the output effect causes a decrease in teaching, while the substitution effect causes an increase. Thus, the change in output is indeterminate.

In the case of a highly accurate screening mechanism with small measurement error, conclusions reached in my earlier work still hold for changes in W_t. Given the normality assumption, changes in W_t will yield like changes in Q_t regardless of the professor's place in the output distribution. In the case of low screening intensity where the measurement error is large, the effect of a change in the income weight given to Q_t is unpredictable. The implications of this are quite clear. For example, if central administration says teaching is now to receive a higher weight in salary allocation, while they do not have an accurate system to measure this output, there may be no change in teaching output resulting; there may be no change in faculty time

[11]One might argue that this analysis ignores the "market" for given faculty skills. However, as cited earlier, the academic labor market, relative to industry or government, is insensitive to market prices for given faculty services. Furthermore, a professor's teaching output contribution is at best known by the employing institution. There is no widely used national information source or market for faculty teaching output. In the case of research, a high producing professor's work is widely known During "research boom periods," such as the 1960's, or within "research boom areas," such as the natural sciences in the late 1950's, a university has to condition its policies on the marketability of high research producers who are being screened by the market itself. During "slack periods" or within "slack areas" even the marketability of the high research producers is restricted.

given to teaching. In light of this finding Tuckman's statement, "If teaching were to receive a greater reward, the means to discriminating among faculty might become more refined [and instructional output might improve]" (p. 119), needs to be reworded. It should read, "If teaching (or any other academic output) were to receive a greater reward, the means of discriminating among faculty must be refined before increased teaching (or any other academic output) can be expected."

IV. Policy Implications

Interestingly, the above theoretical model of professorial behavior provides quite conclusive policy recommendations. In particular, if a given academic output distribution yields a modal output value below the median output, an increase in screening will yield an increase in the given output for the majority of faculty. Furthermore, this increased screening will probably be well received by most faculty, as all those above the mode can expect to gain income from the increased accountability.[12]

On the other hand, if the modal output is above the median output, an increase in screening may or may not result in an increase in the output for the majority of faculty. However, more screening will reduce the variance in measurement in salary determination. Therefore, an appropriate increase in screening combined with subsequent increases in the income determination weight given to the output will raise the level of output which every faculty member is willing and able to produce. Without the matching increase in the income determination weight, however, the increase in screening will probably not be well received by the majority of faculty as they are below the mode and will thus expect to lose income.

In the case of research, the existing screening methods are generally agreed upon and considered highly accurate. For example, an article in a prestigious, refereed journal is accepted as an indicator of quality output. As such, a university interested in raising research output may only need to raise the income determination weight given to research.

Teaching, unlike research output, has no existing measure which is universally accepted as highly accurate. An increase in the income determination weight given this output need not result in an increase in this output. Unless accurate methods of screening are established, there may be no way other than stipulating minimal time commitment regulations to affect this output. Only if a university is able to adopt student evaluations, standardized student learning measures, or some other proxy index of teaching output can the reward structure be used to cause an increase in every faculty member's desire to increase productivity in teaching.

Finally, the introduction of a faculty union which generates pressures for less "invidious comparisons" and more "homogenization" of various subdivisions of the university may produce parity of faculty pay. However, it will also cause those above the modal output levels to reduce their productivity as the recognition of their individual output is mitigated. Similarly, the application of the Equal Pay Act will tend to reduce the performance of the high producers. Any university action resulting in reduced academic screening will have a direct negative effect on the output of high producers. More importantly, reduced screening renders impotent the university's power to increase faculty output through changes in the income weighting given to teaching and research.

[12]Whether an increase in screening is well received depends on the initial screening accuracy. In the extreme case, where no screening is employed, the median person receives the average income. If the mean is above the median, gross income may first fall and then rise as the screening becomes more accurate, see Stiglitz, p. 297.

REFERENCES

W. E. Becker, Jr., "The University Professor as a Utility Maximizer and Producer of Learning, Research and Income," *J. Hum. Resources*, Winter 1975, *10*, 109–15.

Martin Bronfenbrenner, *Income Distribution Theory*, Chicago 1971.

Kenneth E. Eble, *Professors as Teachers*, San Francisco 1972.

Michael K. Evans, *Macroeconomic Activity: Theory, Forecasting, and Control*, New York 1969.

R. B. Freeman, "Demand for Labor in a Nonprofit Market: University Faculty," in Donald S. Humermesh, ed., *Labor in the Public and Nonprofit Sector*, Princeton 1975, 85–133.

Richard A. Lester, *Antibias Regulation of Universities: Faculty Problems and Their Solution*, New York 1974.

Richard B. McKenzie, "The Economics of Reduced Faculty Teaching Loads," *J. Polit. Econ.*, May/June 1972, *80*, 617–19.

D. Needham, "The Economics of Reduced Faculty Teaching Loads: Comment," *J. Polit. Econ.*, Feb. 1975, *83*, 219–23.

J. E. Stiglitz, "The Theory of 'Screening,' Education, and the Distriubution of Income," *Amer. Econ. Rev.*, June 1975, *65*, 283–300.

Howard P. Tuckman, *Publication, Teaching, and the Academic Reward Structure*, Lexington 1976.

NOTES

The ninety-second annual meeting of the American Economic Association will be held in Atlanta, Georgia, December 28–30, 1979. The Professional Placement Service will be open December 27–30.

The Professional Placement Service at the 1979 annual meetings of the Allied Social Science Associations in Atlanta will begin operation on December 27, the day before sessions begin. Applicants and employers will be able to attend more sessions with a day set aside entirely for labor market transactions. This service will be located at the Convention Placement Center in the Marriott Hotel. It will be open from 10:00 A.M. to 5:00 P.M., December 27, 9:00 A.M. to 5:00 P.M., December 28–29; and 9:00 A.M. to 12:00 noon, December 30

Because the 1980 annual meeting comes at an early date in the academic year (September 5–7), the Executive Committee of the American Economic Association has decided to provide employment services at a later time. The 1980 Professional Placement Service will be held at the Dallas Hilton, Dallas, Texas, December 28–30, 1980

Call for Papers for the 1980 Meetings

Members wishing to give papers or make suggestions for the program for the meetings to be held in Denver, September 5-7, 1980, are invited to send their ideas and abstracts to C. Elton Hinshaw, Secretary AEA, 1313 21st Avenue South, Nashville, TN 37212. Although most of the sessions sponsored by the American Economic Association will consist of invited papers, there will also be several sessions of noneconometric contributed papers. (The sessions of contributed papers will not be published in the *Papers and Proceedings* issue to appear May 1981.) Proposals for invited sessions should be submitted as soon as possible. To be considered for the contributed sessions, abstracts of proposed (noneconometric) papers must be received no later than January 15, 1980. Economists wishing to give papers on econometrics or economic theory may submit abstracts to the Econometric Society, which meets with the American Economic Association and annually schedules a substantial number of contributions.

Economists who are strongly oriented toward the humanities, who use humanistic methods in their research, and who will be participating in meetings held outside the United States, Mexico, and Canada that are concerned with the humanistic aspects of their discipline are eligible to apply for small travel grants of the American Council of Learned Societies. Financial assistance is limited to air fare between major commercial airports and will not exceed one-half of projected economy-class fare. Social scientists and legal scholars who specialize in the history or philosophy of their disciplines are eligible if the meeting they wish to attend is so oriented. Applicants must hold a Ph.D. degree or its equivalent, and must be citizens or permanent residents of the United States. To be eligible, proposed meetings must be broadly international in sponsorship or participation, or both. The deadlines for applications to be received in the ACLS office are: meetings scheduled between July and October, March 1; for meetings scheduled between November and February, July 1, for meetings scheduled between March and June, November 1 Please request application forms by writing directly to the ACLS (Attention. Travel Grant Program), 345 East 46th St., New York, NY 10017, setting forth the name, dates, place, and sponsorship of the meeting, as well as a brief statement describing the nature of your proposed role in the meeting Even when plans are incomplete, a prospective applicant should request forms in advance of the cut-off date, since deadlines are firm and no exceptions are permitted Awards will be announced approximately two months after each deadline

---

Omicron Delta Epsilon, the International Honor Society in Economics, announces the winners of the tenth annual Irving Fisher and Frank W Taussig Awards The Final Selection Board (John S. Chipman, Dale W Jorgenson, James A Mirrless, Arnold Zellner, and Egon Neuberger, *ex officio*) declared for the Fisher Award Kenneth J. Singleton, University of Virginia, "The Cyclical Behavior of the Term Structure of Interest Rates", Honorable Mention went to John M. Abowd, University of Chicago, "An Econometric Model of the U.S Market for Higher Education." For the Frank W. Taussig Award: Jeff Eisenach, Claremont Men's College, "The Monetary Policy Debate. Theory and Evidence"; Honorable Mention went to Handen Silver III, Davidson College, "An Economic Justification for Zoning"

Deadlines for submitting entries are February 1, 1980 for the Fisher Competition and June 15, 1980 for the Taussig Competition. Send inquiries and submissions to Egon Newberger, Editor (Tapan Mitra, Associate Editor), Department of Economics, State University of New York, Stony Brook, NY 11794.

Call for Papers. The twelfth annual meeting of CHEIRON: The International Society for the History of Behavioral and Social Sciences will be held at Bowdoin College, June 19-21, 1980. Papers are solicited which deal with topics in the history of any of the behavioral or social sciences, with particular interest in interdisciplinary, methodological, and speculative contributions Deadline for receipt of papers is January 15, 1980. All

submissions must be the complete paper, no longer than eight double-spaced typewritten pages, to be subjected to anonymous refereeing Send papers and queries to Rand Evans, Program Chair, Department of Psychology, Texas A&M University, College Station, TX 77840

There will be a meeting of the Southwestern Association for Slavic Studies in Houston, Texas, April 2 5, 1980. There will also be a combined meeting of the Rocky Mountain and Southwestern Associations for Slavic Studies in Albuquerque, New Mexico, April 23–26, 1980.

In 1975 the Congress established the Alan T Waterman Award to honor the first director of the National Science Foundation This annual award recognizes an outstanding young scientist in the forefront of science. In addition to a medal and other recognition, the recipient receives a grant of up to $50,000 per year for a period of up to three years for scientific research or advanced study in the mathematical, physical, medical, biological, engineering, social, or other sciences at the institution of the recipient's choice The Award Committee annually solicits nominations from a wide variety of sources and invites participation in the nomination process Candidates must be U S citizens and must be 35 years of age, or younger Please submit six copies of the supporting documents (letter size and unbound), including the following complete biography, giving date of birth, detailing the nominee's academic work and career since obtaining the last degree Particular attention should be given to documenting the significance and originality of the nominee's contribution to the advancement of science Also give a one- or two-sentence citation that summarizes the nominee's career and potential for accomplishment Names, complete addresses, and telephone numbers of at least three references who can attest to the nominee's work and career to date and who are located outside the nominee's home institution are to be given The nominations with supporting documentation should be submitted to the Alan T Waterman Award Committee, NSF, Washington, D C 20550 (No special forms are required.)

The 1978 winner of the Leonard J Savage Thesis Award for an outstanding thesis in Bayesian econometrics and statistics was Lorraine De Robertis Her doctoral dissertation, "The Use of Partial Prior Knowledge in Bayesian Inference," was completed at Yale University Honorable Mention was accorded to José M Bernardo for his thesis, "The Use of Information in the Design and Analysis of Scientific Experimentation," completed at University College, London. The Thesis Evaluation Committee included R E. Kihlstrom, chairman, C. A Barry, E. E. Leamer, R A Olshen, J W Pratt, and A. Zellner, *ex officio*. Entries for the 1979 Savage Award are currently being judged.

The Carolina Population Center invites applications for a small number of postdoctoral fellowships for population research training in sociology, psychology, economics, anthropology, epidemiology, and maternal and child health. Postdoctoral programs of the following types will be considered 1) Specialized training in the discipline of the applicant's doctorate for those holding doctorates involving population research training; 2) Training in a second discipline for those with doctorates involving population research training; 3) Individualized interdisciplinary programs for those with doctorates which may (but need not) involve population research training. Applicants must be U.S. citizens or permanent residents of the United States There are no application deadlines. Direct inquiries to J Richard Uldry, Director, Carolina Population Center, University of North Carolina, Chapel Hill, NC 27515

The S S Huebner Foundation for Insurance Education is sponsoring its fifth annual research grants competition The grants are intended to support research in the field of risk and insurance Full-time faculty members at colleges and universities in the United States and Canada are eligible to apply for grants Applicants should hold a terminal degree such as Ph D or D.B.A , a law degree, or be a fellow of an actuarial society. Among the general topic areas which will be considered for grants are risk theory, consumer demand for insurance, risk management, health insurance, other methods of health care financing, social insurance programs, international insurance issues, insurance law, and insurance regulation. Grants are available from the Foundation in amounts up to $10,000 Proposals must be submitted by March 1, 1980, and the grants will be awarded by June 1, 1980. Additional information regarding the program can be obtained from Dr J David Cummins, Research Director, S S Huebner Foundation for Insurance Education, Colonial Penn Center, 3641 Locust Walk CE, Philadelphia, PA 19104

The Women's Educational Equity Act Program, under the auspices of the Office of Education of the U.S. Department of Health, Education, and Welfare, is a federally funded program designed to promote educational equity for women in the United States. As a part of this national goal, AMERICAS Behavioral Research Corporation was awarded a grant to develop The Center for Women Scholars (CFWS). The purpose of CFWS is to address the causes and effects of institutionalized discrimination against women scholars.

The CFWS is offering a $500 prize for the best article of not more than 5,000 words on solutions to the problems of the woman scholar, and publication in *The Woman Scholar's Handbook: Strategies for Success*, forthcoming 1980 Address inquiries to Dr. Mary Spencer, and send submissions to Dr. Monika Kehoe, Editor, CFWS, AMERICAS, 300 Broadway, Suite 23, San Francisco, CA 94133

Economics of Education Review, a new journal published by Ballinger Publishing Company, invites manuscripts concerning all facets of the economics of education. The *EER* is devoted to the development of sound theoretical, empirical, and policy research in the economics of education, demonstrating the potential role of economic analysis in the solution or improved understanding of educational problems and issues. Write to Elchanan Cohn, Editor, Department of Economics, College of Business Administration, University of South Carolina, Columbia, SC 29208.

Proceedings now available (free) upon request for two National Science Foundation 1977 conferences. "Economic and Demographic Methods for Projecting Population" and "Development of User Oriented Software." For single copies of either or both, write American Statistical Association, 806 15th St., NW, #640, Washington, D.C. 20005.

Call for Papers: The Department of Philosophy of Georgia State University will hold an Interdisciplinary Conference on Capital Punishment in Atlanta, April 18–19, 1980. Academic fields represented will include philosophy, religion, sociology, criminology, law, psychology, political science, and economics The deadline for submitted papers is December 15, 1979. Selections will be announced February 1, 1980. Please send papers and inquiries to Professor C. G. Luckhardt, Department of Philosophy, Georgia State University, University Plaza, Atlanta, GA 30303.

Deaths

Otto H. Ehrlich, professor emeritus of economics, Graduate School of Arts and Science, New York University, and founder and editor of *Economic Abstracts*, May 22, 1979.

Donald S. Tucker, emeritus professor of economics, M.I T , Feb. 16, 1979

Retirements

Nathan M. Becker: professor emeritus in residence, Pace University, Sept 1979

Carl D. Snyder, professor of economics, Eastern Michigan University, June 30, 1979.

Visiting Foreign Scholars

Zvi Adar, Tel Aviv University, Israel: visiting professor, department of economics, University of California-Davis, July 1, 1980.

Wilford Beckerman, Balliol College, England: visiting alumni professor, department of economics, Clemson University, Jan.-May 1980.

Philip Black, Rhodes University, South Africa: visiting associate professor of economics, Pennsylvania State University, Sept. 1, 1979.

Francois Bourguignon, École Normale Superieure, France: visiting scholar, department of economics, Concordia University, Jan.-May 1980.

Kenneth A. Chrystal, University of Essex, England: visiting assistant professor, department of economics, University of California-Davis, July 1, 1979.

Raymond DeBondt, Katholieke Universitet te Leuven, Belgium: visiting associate professor, department of managerial economics and decision sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979-Mar. 15, 1980.

Gerald C. Huber, University of Geneva, Switzerland: visiting scholar, department of economics, M.I T., Aug 16-May 31, 1980

Yoshitaka Itsumi, Gakushuin University, Japan. visiting scholar, department of economics, M.I T., Aug. 28, 1979-Aug. 27, 1980

Peter Lambert, University of York, England: visiting professor, department of economics, Lehigh University, Spring 1980.

Richard G Lipsey, Queen's University: visiting professor of economics, Yale University, Sept. 1979.

James P. Neary, Trinity College, Ireland: visiting scholar, department of economics, M.I.T , July 1, 1979-Sept. 30, 1979.

Kevin W S Roberts, Oxford University, England visiting assistant professor, department of economics, M.I.T , July 1, 1979-Aug. 31, 1979.

Brinley Thomas, University of Cardiff, Wales: visiting professor, department of economics, University of California-Davis, Jan 1, 1980

Dror Zuckerman, Hebrew University of Jerusalem visiting assistant professor, department of managerial economics and decision sciences, J L Kellogg Graduate School of Management, Northwestern University, Sept 1, 1979-Aug 31, 1980

Promotions

Jean W Adams associate professor of economics, Iowa State University, Sept 1, 1979

Jaleel Ahmad: professor of economics, Concordia University, June 1978.

Ryan C Amacher professor of economics, Arizona State University, Aug. 25, 1979.

Donald V. T Bear: professor or economics, University of California-San Diego, July 1, 1979

William C Bonifield: professor of economics, Wabash College, July 1, 1979

John B Burbidge: associate professor of economics, McMaster University, July 1979.

Vittorio Corbo: professor of economics, Concordia University, June 1, 1979.

Karl Case: associate professor, economics department, Wellesley College, Sept. 1980.

Cheryl Casper. associate professor of economics, Kent State University, July 1, 1980.

Charles D. Delorme: professor, department of economics, University of Georgia, Sept. 1978.

Rudiger Dornbusch: professor of economics, M.I T., July 1978.

Louis H. Ederington: professor of economics, Georgia State University, Sept. 1979.

Walter Enders: associate professor of economics, Iowa State University, Sept. 1, 1979.

Ray C. Fair: professor of economics, Yale University, July 1, 1979.

William A. Fischel: associate professor, department of economics, Dartmouth College, July 1, 1979.

Roy G. Gardner: associate professor of economics, Iowa State University, Sept. 1, 1979

Robin Hahnel: assistant professor of economics, American University, June 1, 1979.

Jerry A. Hausman. professor of economics, M I.T., July 1, 1979.

N. Nurul Islam: associate professor, department of economics, Concordia University, June 1, 1979

Marvin R. Jackson. professor of economics, Arizona State University, Aug. 25, 1979

Paul L Joskow: professor of economics, M.I T , July 1978

Hiroaki Kayakawa associate professor, department of economics, University of Georgia, Sept. 1979.

Steven W Kohlhagen. associate professor, School of Business Adminstration, University of California-Berkeley, July 1, 1979.

Jerry R Ladman professor of economics, Arizona State University, Aug. 25, 1979.

Harvey E Lapan. professor of economics, Iowa State University, Sept. 1, 1979.

Nicholas R. Lardy associate professor of economics, Yale University, July 1, 1979

Richard C. Levin associate professor of economics, Yale University, July 1, 1979

Gerald S. McDougall: associate professor, department of economics, Wichita State University University, Aug 1979

Roger Myerson associate professor, department of managerial economics and decision sciences, J L Kellogg Graduate School of Management, Northwestern University, Sept 1, 1979

R Andrew Muller. associate professor of economics, McMaster University, July 1979.

Narayan K. Nargund: professor of economics, Allegheny College, July 1979.

Douglas Needham: professor of economics, Western Kentucky University, Aug 1979

Sharon M. Oster. associate professor of economics, Yale University, July 1, 1979

Nallapu N. Reddy: professor of economics, University of Michigan-Flint, Sept. 1, 1979.

Charles A. Roberts. associate professor of economics, Western Kentucky University, Aug. 1979.

Mark E. Schaefer: associate professor of economics, Georgia State University, Sept. 1979.

Parthasarathi Shome: associate professor of economics, American University, July 1, 1979.

Barton Smith: associate professor of economics, University of Houston, Sept. 1, 1979.

Kenneth I. Wolpin. associate professor of economics, Yale University, July 1, 1979.

M Raquibuz Zaman. associate professor, School of Business, Ithaca College, May 1979.

Eitan Zemel: associate professor, department of managerial economics and decision sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979.

Administrative Appointments

Carolyn Shaw Bell: chairman, economics department, Wellesley College, July 1979.

Robert W. Campbell: chairman, department of economics, Indiana University, July 1978.

Cheryl Casper: assistant dean, College of Business Administration, Kent State University, July 1, 1979.

George Daly: dean, College of Social Sciences, University of Houston, Sept. 1, 1979.

Edward T Dowling: chairman, department of economics, Fordham University, Sept. 1, 1979.

Elmer R. Gooding: assistant provost, Arizona State University, July 1979.

Samuel Gubins. vice president for planning, and treasurer, Haverford College, July 1979

E Yong Lee chairman, department of accounting, business and economics, Upsala College, July 1, 1979.

Robert Maurer. chairman, department of economics and business management, Central College, Sept. 1979.

Narayan K Nargund chairman, department of economics, Allegheny College, July 1979

Monroe Newman. head, department of economics, Pennsylvania State University, Apr. 1, 1979.

Albert W. Niemi: director of research, College of Business Administration, University of Georgia, Jan. 1979

John W. Rowe, University of Houston. head, department of economics, University of South Florida, Sept. 1, 1979

Balbir S. Sahni. chairman, department of economics, Concordia University, Jan 1978.

Mark Satterthwaite. chairman, department of managerial economics and decision sciences, J L Kellogg Graduate School of Management, Northwestern University, Sept 1, 1979.

G Edward Schuh, Purdue University chairman, department of agricultural economics, and professor of economics, University of Minnesota, Sept. 1979.

K. Dickson Smith, Lakeland College: chairperson, and associate professor, department of business, Cardinal Stritch College, Sept 1, 1979.

M. Raquibuz Zaman: chairman, department of accounting and finance, School of Business, Ithaca College, Aug 1979.

Appointments

Badi Baltagi, University of Pennsylvania. assistant professor of economics, University of Houston, Sept. 1, 1979

Sandra Baum. instructor, economics department, Wellesley College, Sept. 1979.

William E. Becker, Jr.: associate professor of economics, Indiana University, August 1979.

John J. Beggs, Northwestern University: assistant professor of economics, Yale University, July 1979.

Arthur Blakemore: assistant professor of economics, Arizona State University, Aug. 1979.

Joseph Burns, Commodities Futures Trading Commission: economist, Antitrust Division, U.S. Department of Justice, Apr. 1979.

Stanford Calderwood: visiting professor, economics department, Wellesley College, Sept. 1979.

Philippe Callier, Simon Fraser University: lecturer, department of economics, Concordia University, Jan. 1979.

James D. Christensen: instructor, department of economics, Iowa State University, Mar. 1, 1979.

George Cluff: assistant professor of economics, Georgia State University, Jan. 1979.

John A. Edgren, University of Michigan: assistant professor of economics, Eastern Michigan University, Sept. 1979.

William M. Edwards: assistant professor, department of economics, Iowa State University, Mar. 1, 1979.

David Eisenstadt, University of Missouri: economist, Antitrust Division, U.S. Department of Justice, Aug. 1979.

Larry Eubanks, University of Wyoming: assistant professor of economics, Pennsylvania State University, Sept. 1, 1979.

Joseph v.R. Farrell, Oxford University: instructor, department of economics, M.I T., Sept. 1979.

Robert Ford, University of Calgary: lecturer, department of economics, Concordia University, June 1979.

David Forrest: visiting lecturer, department of economics, McMaster University, July 1979.

Celeste Gaspari: instructor, economics department, Wellesley College, Sept 1979

John H Gates, College of William and Mary. Bureau of Economic Analysis, U S. Department of Commerce, July 1, 1979.

John Geanakoplos, Harvard University. assistant professor of economics, Yale University, July 1979.

James N. Giordano, Indiana University instructor, department of economics, Fordham University, Sept 1, 1979.

Diego Giurleo: visiting lecturer, department of economics, McMaster University, July 1979.

Lawrence Glosten, Data Resources, Inc.. assistant professor, department of managerial economics and decision sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979.

Robert Gough: visiting lecturer, economics department, Wellesley College, Sept. 1979.

Eric S. Graber: assistant professor of economics, St. Louis University, July 1, 1979.

James Grant: instructor, economics department, Wellesley College, Sept. 1979

Marguerite Guerin-Calvert, Princeton University: economist, Antitrust Division, U.S. Department of Justice, July 1979.

Jeffrey Hammer, M.I.T.: department of economics, University of California-San Diego, July 1, 1979.

John Haring, Civil Aeronautics Board: economist, Antitrust Division, U.S. Department of Justice, July 1979.

Robert A. Hart: visiting associate professor, department of economics, McMaster University, Sept. 1979.

James Hartigan, North Carolina State University: assistant professor, Pennsylvania State University, Sept. 1, 1979.

James H. Hodge: economist, Business Conditions Division, domestic research department, Federal Reserve Bank of New York, July 9, 1979.

James Hoehn, University of Virginia: instructor in economics, Pennsylvania State University, Sept. 1, 1979

Dennis Hoffman: assistant professor of economics, Arizona State University, Aug. 1979.

Bengt Holmström, Swedish School of Business Administration: assistant professor, department of managerial economics and decision sciences, J. L Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979.

Janet C. Hunt: assistant professor, department of economics, University of Georgia, Sept. 1978.

Judith Hushbeck: instructor of economics, American University, Sept. 1, 1979.

Ahmed Hussein: assistant professor of economics, Allegheny College, Sept 1979

Farrukh Iqbal: instructor, economics department, Wellesley College, Sept. 1979.

Ian Irvine, University of Western Ontario: assistant professor, department of economics, Concordia University, June 1978

Takatoshi Ito, National Bureau of Economic Research, and Harvard University: assistant professor of economics, University of Minnesota, Sept 1979

Antony Jackson, Bristol University, England: assistant professor, department of economics, Concordia University, June 1979

Janet Johnson, University of Wisconsin: assistant professor department of economics, Syracuse University, Sept. 1, 1979.

Robert W. Jolly: assistant professor, department of economics, Iowa State University, May 1, 1979.

Donald C. Keenan: assistant professor, department of economics, University of Georgia, Sept. 1978.

Elsie M. Knoer, Arizona State University. assistant professor of economics, University of California-Davis, July 1979

Tetteh Kofi, Stanford Food Research Institute: visiting associate professor, department of economics, University of Notre Dame, Fall 1979-80.

Tom K. Lee, California Institute of Technology department of economics, University of California-San Diego, July 1, 1979.

Linda Leighton, Queens College: assistant professor, department of economics, Fordham University, Sept. 1, 1979.

Lucinda M. Lewis, University of Pennsylvania: economist, Antitrust Division, U.S. Department of Justice, Sept. 1979.

Robert Litterman, University of Minnesota: assistant professor, department of economics, M.I.T., Sept. 1979.

Stuart Low: assistant professor of economics, Arizona State University, Aug. 1979.

Carma McClure: assistant professor, department of managerial economics and decision sciences, J. L. Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979.

Daniel L. McFadden, University of California-Berkeley: professor of economics, M.I.T., Sept. 1978.

Andrew McLaughlin, University of California-Los Angeles: economist, Antitrust Division, U.S. Department of Justice, Sept. 1979.

Mark Machina, M.I.T.: department of economics, University of California-San Diego, July 1, 1979.

Thomas Martin, Rice University: instructor, department of economics, Clemson University, Aug. 1979

John A. Miller, University of Pittsburgh: assistant professor, department of economics, Wheaton College, Sept. 1979.

Ross Miller, Harvard University. assistant professor of economics, University of Houston, Sept. 1, 1979

Lazaros E. Molho, Yale University. instructor, department of economics, Fordham University, Sept 1, 1979

Paul A. Montavon. director, USAID mission to Paraguay, Asuncion, June 22, 1979

Hannah M. O'Connell. visiting associate professor, American University, Sept 1, 1979.

Mary O'Keeffe, Harvard University· visiting assistant professor of economics, University of Houston, Sept 1, 1979.

Chris W Paul assistant professor, department of economics, University of Georgia, Sept 1977

Mary K. Perkins· instructor of economics, American University, Sept 1, 1979

Peter C B. Phillips, University of Birmingham, England. professor of economics, Yale University, July 1, 1979

Richard A. Phillips, University of North Carolina assistant professor of economics, Pennsylvania State University, Sept 1, 1979

Michael Podgursky assistant professor of economics, University of Notre Dame, Fall 1979-80

Joseph M Pogodzinski visiting assistant professor of economics, Georgia State University, Sept 1979

Donald Poskitt, University of York, England visiting professor, department of economics, Lehigh University, Fall 1979.

Michael D Pratt assistant professor of economics, Virginia Commonwealth University, Aug 1979

Francisco L. Rivera-Batiz· lecturer, department of economics, Indiana University, Aug. 1979

Jennifer Roback, University of Rochester assistant professor of economics, Yale University, July 1979.

Bruce B Roberts assistant professor, College of William and Mary, Sept 1979.

Mark Rosenzweig, Yale University associate professor of economics, University of Minnesota, Sept. 1980

Larry Samuelson, University of Florida assistant professor, department of economics, Syracuse University, Sept. 1, 1979.

John Sawyer, University of Pennsylvania· economist, Antitrust Division, U.S. Department of Justice, June 1979.

Willy Sellekaerts: visiting professor of economics, American University, Sept. 1, 1979

Willi Semmler: assistant professor of economics, American University, Sept 1, 1979.

Eckhard Siggel, University of Toronto· assistant professor, department of economics, Concordia University, June 1978.

Christopher Sims, University of Minnesota. visiting professor of economics, M.I.T., Sept. 1979.

Joel Slemrod, Harvard University: assistant professor of economics, University of Minnesota, Sept. 1979.

David Smith, University of Michigan. economist, Antitrust Division, U.S. Department of Justice, July 1979.

J. Barry Smith, University of Western Ontario: assistant professor, department of economics, Concordia University, June 1978.

T. N. Srinivasan, the World Bank: professor of economics, Yale University, July 1, 1979.

Jerry L Stevens· assistant professor of business economics, Colgate Darden Graduate School of Business Administration, University of Virginia, Sept. 1979.

Lawrence Summers, Harvard University: assistant professor of economics, M I T., Sept 1979

Caroline Swartz: assistant professor, department of economics, University of Notre Dame, 1979-80

Thomas Teisberg, U.S. Department of the Interior. assistant professor of economics, M.I.T., Sept. 1978.

D Rodney Thom. visiting assistant professor, department of economics, McMaster University, Sept. 1979.

Richard Trainer economist, communications development staff, public information department, Federal Reserve Bank of New York, May 14, 1979.

Claude M Vaughan, Kentucky Development Cabinet: commissioner, Energy Regulatory Commission, Apr 1, 1979

Michael R Veall· visiting lecturer, department of economics, McMaster University, July 1979.

Robert Weber, Yale University associate professor, department of managerial economics and decision sciences, J L Kellogg Graduate School of Management, Northwestern University, Sept. 1, 1979.

Rona Weiss instructor, economics department, Wellesley College, Sept 1979

David E Wildasin· assistant professor of economics, Indiana University, Aug 1979

Arlington Williams: assistant professor of economics, Indiana University, Aug. 1979.

John Willoughby visiting assistant professor of economics, American University, Sept 1, 1979

Stephen Woodbury, University of Wisconsin: assistant professor of economics, Pennsylvania State University, Sept. 1, 1979

Leaves for Special Appointment

Lloyd Atkinson, American University: senior economist, Joint Economic Committee, U.S. Congress.

Nancy S Barrett, American University Deputy Assistant Secretary of Labor, U.S. Department of Labor.

John M Clapp, University of California-Los Angeles: economist, Program Analysis Division, General Accounting Office, Sept 15, 1979-Sept. 1980.

Thomas F. Dernburg, American University: staff director, Subcommittee on Economic Stabilization of the Senate Committee on Banking Finance and Urban Affairs, U.S. Senate

Alan DeSerpa, Arizona State University: visiting professor, Naval Post Graduate School, Aug. 1979.

Jonathan Dickinson, Pennsylvania State University: economist, Stanford Research Institute, Dec. 1, 1979-Dec. 1, 1980

Louis H. Ederington, Georgia State University: Fulbright lectureship, Academy of Economic Studies, Bucharest, Jan.-June 1980.

Robert M. Feinberg, Pennsylvania State University:

economist, Antitrust Division, U.S. Department of Justice, Aug. 1, 1979-June 30, 1980.

Klaus Friedrich, Pennsylvania State University: economist, Division of International Finance, Board of Governors of the Federal Reserve System, June 30, 1979-June 30, 1980.

Joanna Frodin, Wellesley College: Citibank, New York.

R. Jeffery Green, Indiana University: director, quarterly model project, Wharton Econometric Forecasting Associates, May 1979-Aug. 1980.

Hiroaki Hayakawa, University of Georgia: Nanzan University, Japan, June 1979.

Charles Holt, University of Minnesota: visiting professor of economics, University of Virginia, 1979-80.

Marvin R. Jackson, Arizona State University. Institute for World Economy, Bucharest.

Hiromitsu Kaneda, University of California-Davis. International Institute for Applied Systems Analysis, Vienna, Fall 1979.

Kwan S. Kim, University of Notre Dame visiting professor, Agency for International Development, 1979-81.

Thomas A. Layman, Arizona State University. economist, Crocker National Bank, Aug 1979

John S. Lyons, American University temporary lecturer, department of economics, University of Essex, England.

Robert Machol, Northwestern University: Office of Naval Research, London, 1978-80.

Mark Pitt, University of Minnesota. Ford Foundation Consultant to Planning Cell, Ministry of Agriculture, Bangladesh, 1979-80.

Mark Rosenzweig, Yale University. Population Center, University of Michigan, 1979-80.

Robert J. Thornton, Lehigh University: visiting research fellow, University of Sussex, Fall 1979.

George M von Furstenberg, Indiana University: economist, International Monetary Fund, Aug. 1979-May 1980.

James N. Wetzel, Virginia Commonwealth University· Boston University Overseas Program, 1979-80.

Mahmood A. Zaidi, University of Minnesota: visiting professor, University of New South Wales, Australia, 1979-80.

Resignations

James D. Adams, Iowa State University: University of Florida, May 31, 1979.

Martin N. Baily, Yale University: The Brookings Institution, Aug. 1979.

M. E. Bond, Arizona State University, May 15, 1979.

Michael Carter, University of Notre Dame, Sept. 1, 1979.

Susan B Carter, University of Notre Dame, Sept. 1, 1979.

George Cochran, Kent State University, June 8, 1979

Nancy Griffeth, Northwestern University: Georgia Institute of Technology, Sept. 1, 1979.

Christopher J. Heady, Yale University. University College, London, June 1979.

Stephen Horner, Wellesley College, July 1979.

Dean Kropp, Northwestern University. Dartmouth College, Sept. 1, 1979.

James M. McGrann, Iowa State University: Texas A&M University, Mar. 31, 1979.

Richard Pomfret, Concordia University, June 1979.

David H. Resler, St. Louis University Federal Reserve Bank of St. Louis, July 1, 1979.

James S. Sagner, Southern Illinois University-Edwardsville: A. T. Kearney, Chicago, Sept. 15, 1979.

Edward Stohr, Northwestern University: New York University, Sept 1, 1979.

William Weiskopf, Kent State University, Jan. 1, 1980.

NOTE TO DEPARTMENTAL SECRETARIES AND EXECUTIVE OFFICERS

When sending information to the *Review* for inclusion in the Notes Section, please use the following style:

A. Please use the following categories·

1—Deaths
2—Retirements
3—Foreign Scholars (visiting the USA or Canada)
4—Promotions
5—Administrative Appointments
6—New Appointments
7—Leaves for Special Appointments (NOT Sabbaticals)
8—Resignations
9—Miscellaneous

B. Please give the name of the individual (SMITH, Jane W), her present place of employment or enrollment: her new title (if any), and the date at which the change will occur.

C. Type each item on a separate 3 x 5 card and please do not send public relations releases.

D. The closing dates for each issue are as follows: *March*, October 15; *June*, January 15; *September*, April 15, *December*, July 15.

This announcement supersedes and replaces a letter which was sent annually from the managing editor's office. All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

SEVENTY-SIXTH LIST OF DOCTORAL DISSERTATIONS IN POLITICAL ECONOMY IN AMERICAN UNIVERSITIES AND COLLEGES

The present list specifies doctoral degrees conferred during the academic year terminating June 1979. Abstracts will no longer be printed, as they are published by University Microfilms, Ann Arbor, Michigan.

General Economics; including Economic Theory, History of Thought, Methodology, Economic History, and Economic Systems

EDMUNDO ACOSTA, Ph.D. Washington (St Louis) 1978. Two monetary approaches to the economic history of the United States, 1842–59.

DANIEL R. ALGER, Ph.D. Northwestern 1979 Markets where firms select both prices and quantities. An essay on the foundations of micro-economic theory.

BETH E. ALLEN, Ph.D. California (Berkeley) 1978 Generic existence of equilibria for economies with uncertainty when prices convey information

M. RAHIM J. AMIN, Ph.D Cincinnati 1978 Paradigm shift in economic science.

KAMAL ARGHEYD, D B.A. Harvard 1978 The role of value systems in the process of social change. The Shah-people revolution of Iran

ROGER A ARNOLD, Ph.D. Virginia Polytechnic Institute 1979. The efficiency properties of institutional evolution with particular relevance to the social-philosophical works of F. A. Hayek

SHERRY S ATKINSON, Ph.D. Pittsburgh 1978 Microtheoretic analysis of inventory behavior

PIERRE BATTEAU, Ph D. Northwestern 1979 Essays in the theory of collective choice

RANDALL S BAUSOR, Ph.D Duke 1978 A general equilibrium model of perceptions, expectations, strategies, and outcomes

BEN S. BERNANKE, Ph D Massachusetts Institute of Technology 1979. Long-term commitments, dynamic optimization, and business cycles

PETER C BING, Ph.D. Massachusetts Institute of Technology 1977. Advertising and economic welfare

BRIAN R. BINGER, Ph.D California Institute of Technology 1979 Essays in forward markets and the uranium industry.

R. MICHAEL BLACK, Ph D California (San Diego) 1979 Two economic applications of dynamic optimization theory.

MONGI BOUGHZALA, Ph.D. Minnesota 1978 Prices vs quantities as instruments of plan implementation

GEORGE M. BOYD, Ph.D. Florida State 1979 Testing the Ricardian trade model· An economic-historical approach.

EDWARD L. BUBNYS, Ph.D. Illinois 1978. Chicago, 1870 and 1900: Wealth, occupation, and education.

BRUCE J. CALDWELL, Ph.D. North Carolina (Chapel Hill) 1979. Recent methodological thought in economics: Survey and critique from a philosophy of science perspective.

JAMES D. CAMPBELL, Ph.D. Minnesota 1979. Existence of approximate competitive equilibrium in a nonconvex economy with fiat money.

RAMÓN CAO-GARCIA, Ph.D. Virginia Polytechnic Institute 1979. Explorations toward an economic theory of political systems.

ELIANA A CARDOSO, Ph.D. Massachusetts Institute of Technology 1979. Inflation, growth, and the real exchange rate. Essays on economic history in Brazil.

CLARK J. CHANDLER, Ph.D. Michigan 1978. The effects of the private sector on the labor behavior of Soviet collective farmers.

AJIT K CHAUDHURY, Ph D Rochester 1979. Essays on economic policy under non-market-clearing conditions

SHIH-MENG S. CHEN, Ph D. Ohio State 1978. Theory of quantity-setting firm and risk aversion: A certainty-equivalent approach

EUN KWAN CHOI, Ph D Iowa 1978. Riskiness and risk aversion with many commodities

EDWARD H. CLARKE, Ph D Chicago 1978. Demand revelation and public goods

ROGER P CONGLETON, Ph D Virginia Polytechnic Institute 1978 The role of information and choice.

MICHAEL A. CONTE, Ph.D Michigan 1979. On the economic theory of the labor-managed firm under socialism

LARRY R CROSS, Ph.D Colorado 1979. A theoretical and empirical investigation of some social choice processes

DAVID A. CROWE, Ph.D. Kentucky 1978. Commuter congestion charges and efficient land values.

JOHN F. CUSHMAN, Ph.D. Virginia 1979. Charitable giving and philanthropic foundations: An economic analysis

RUSSELL DAVIDSON, Ph.D British Columbia 1977. Essays in comparative dynamics.

JOSE M A. G. DE CARMARGO, Ph.D. Massachusetts Institute of Technology 1978. The role of labor in the transition to capitalism: The case of the coffee plantations in Sao Paulo, Brazil, 1880–1925.

ERICK C. DUESING, Ph.D. North Carolina (Chapel Hill) 1978. Polyhedral convex sets and the economic analysis of production.

DAVID A. EASLEY, Ph.D. Northwestern 1979. The foundation of expectations in an uncertain environment.

JOHN A. EDGREN, Ph.D. Michigan 1979. Economic behavior of individuals.

CARLOS S. ESCRIBANO, Ph.D. Minnesota 1978. Resource allocation mechanisms yielding Walrasian allocations at equilibrium messages.

PHILLIP E. FINCHER, Ph.D. Mississippi 1978. Some economic welfare implications of a national minimum wage law as a national economic policy.

NANCY R. FOLBRE, Ph.D. Massachusetts 1979. Patriarchy and capitalism in New England, 1620 1800

DAVID I. FOLKERTS-LANDAU, Ph.D. Princeton 1979. Intertemporal planning, exchange, and macroeconomics.

DONALD H. FULLERTON, Ph D California (Berkeley) 1978. A general equilibrium taxation model with applications and dynamic extensions

DAVID W. GALENSON, Ph.D. Harvard 1979. The indenture system and the colonial labor market An economic history of white servitude in British America.

GARY A. GIGLIOTTI, Ph.D. Columbia 1978. Three essays concerning the optimal income distribution, the optimum rate of population growth, and Rawlsian social choice.

JAMES W GILLULA Ph.D Duke 1978 Regional interdependence and economic development in the *USSR*. An interregional input-output analysis

JAMES A. GIROLA, Ph D Harvard 1978 The endogenous business cycle

LAWRENCE GLOSTEN, Ph D Northwestern 1979 Informational efficiency in stock markets with disequilibrium trading

ROBIN HAHNEL, Ph D American 1979. Towards a political economy of socialism

DENNIS HOFFMAN, Ph D. Michigan State 1978 The rational expectations hypothesis A framework for solutions with econometric implications

ELIZABETH HOFFMAN, Ph D. California Institute of Technology 1979 Essays in optimal resource allocation under uncertainty with capacity constraints

HARUO IMAI, Ph.D Stanford 1979 Voting, bargaining, and factor-income distribution

TAKATOSHI ITO, Ph.D. Harvard 1979. Contributions to disequilibrium economic analysis.

MARCUS W. JARSULIC, Ph.D. Pennsylvania 1978. An econometric analysis of enterprise expenditure decisions under Yugoslav self-management

RAGHBENDRA JHA, Ph.D. Columbia 1978. Essays in economic theory

KOSE JOHN, Ph.D. Florida 1978 Computing equilibrium paths of continuous families of economies: A homotopy technique and algorithms.

BENSON KANDOOLE, Ph.D. Southern California 1979. A non-linear simulation model of the American economy.

TIMOTHY KEHOE, Ph.D. Yale 1979. Regularity and index theory for economic equilibrium models.

AKBAR A. KHAN, Ph.D. Queen's (Kingston) 1979 Some aspects of peasant behavior in Bengal, 1890–1914: A neoclassical analysis

NORMAN L. KLEINBERG, Ph.D. Massachusetts Institute of Technology 1978 Three essays in mathematical economics

DAVID H. KNAPP, Ph.D. California (Santa Barbara) 1978. On the instability of spatial equilibria with locational externalities.

ELSIE S KNOER, Ph D Kansas 1978. Economics as probability measures: Results for equilibria

HIROSHI KODAIRA, Ph D Rochester 1979. Majority voting in a stock market economy with uncertainty.

THANONG LAMYAI, Ph D Northwestern 1978 Microanalytic investigation of mechanical engineering production processes

SEUNG-DONG LEE, Ph D Southern Methodist 1979 A dynamic theory of a resources extractive firm

SØREN Q. LEMCHE, Ph D Minnesota 1978. The first optimality theorem with interdependent preferences and public commodities

DAVID M LEVY, Ph.D. Chicago 1979 The content of classical economics

WILLIAM J LYNK, Ph.D Chicago 1978 Information, advertising, and the structure of the market

RICHARD D MACMINN, Ph D Illinois 1977 Search and market equilibrium

SILVANA O MALLE, Ph D California (Berkeley) 1979 The war communism economic organization, 1918 21

PETER A MARKS, Ph D. North Carolina (Chapel Hill) 1978 The use of prices as a planning instrument· A welfare analytic approach with applications

ROBERT E MARTIN, Ph D Southern Methodist 1979 Random production. Stochastic input deliveries and machine failures.

JULIE A MATTHAEI, Ph D. Yale 1978 Womanhood and economic life. A study in American economic history

DANIEL L. MILKOVE, Ph.D. Purdue 1978 Three essays on temporary equilibrium.

ROSS M. MILLER, Ph D Harvard 1979 Signaling, welfare, and optimal taxation

WILLIAM C. MILLER, Ph.D Minnesota 1978 The dynamics of social decision processes with application to an economic exchange model

JEFFREY T. MILLS, Ph.D. Illinois 1979. The economics of price control. The OPA experience, 1941–46.

RENÉ MORENO, Ph.D Northwestern 1978 The distribution of income and economic well-being in the United States, 1968–71

HASSOUNA MOUSSA, Ph.D. Minnesota 1978 The unequal distribution of information and the essentiality of fiat money in general equilibria

HIROAKI NAGATANI, Ph.D. Massachusetts Institute of

Technology 1978 Neoclassical theories of income distribution.

REID NAGLE, Ph.D Johns Hopkins 1979 Consumption and income distribution. The revised standard income theory with application to the Great Depression.

MANIZHEH NAKHAI-AGHMIUNY, Ph D. Southern California 1978. Democratic socialism Reconciling social justice and economic growth

RON W. NAPIER, Ph.D Harvard 1979 Prometheus absorbed: The industrialization of the Japanese economy, 1905–37

WILLIAM NOVSHEK, Ph.D Northwestern 1979. Essays on equilibrium with free entry

WILLIAM W. NYE, Ph.D Stanford 1979 The impact of the American Civil War on the pattern of international trade and on the British money market.

MARK H. OBRINSKY, Ph D Pennsylvania 1978 A survey of profit theory.

NAI-PEW ONG, Ph D Yale 1979 Ricardo and Marx The development of homogeneous physical capital and the advance beyond

MARTHA W PAAS, Ph D Bryn Mawr 1979. Population change, labor supply, and agriculture in Augsburg, 1480–1618 A study of early demographic-economic interactions

SEVKET PAMUK, Ph D California (Berkeley) 1978 Foreign trade, foreign capital, and domestic transformation: The Ottoman Empire in the nineteenth century.

LUCAS D PAPADEMOS, Ph D Massachusetts Institute of Technology 1978 Optimal aggregate employment policy

SPYRIDEN G PATTON, Ph D Pittsburgh 1978 Some impacts of the Reading Railroad on the industrialization of Reading, Pennsylvania, 1838-1910

GENE P PETERSON, Ph.D Colorado State 1979 Private property in Marxian and Catholic social thought

WALKER POLLARD, Ph D Virginia Polytechnic Institute 1978 The effects of economic conditions on elections—A study controlling for political variables

SIMON A RAVID, Ph.D Cornell 1979 Issues in optimal decision making of public utilities under uncertainty

JENNIFER F. REINGANUM, Ph D Northwestern 1979 Dynamic games of *R&D* with rivalry.

DONALD E. RENNER, Ph.D Cincinnati 1979. Approach to changing income distribution

CHARLES A. ROBERTS, Ph.D Georgia 1979. Regional variations in agricultural wage rates, 1870–1920.

JOE O ROGERS, Ph.D. Duke 1978. The impact of economic theory on public policy The case of technological change and employment policy.

LEONARD E RYCHTANEK, Ph D Florida State 1979 In search of a behavioral theory of economics using neurophysiology and anthropology as a foundation

JAHANGIER SALEH, Ph.D. Massachusetts 1978 Social formations in Iran, 750-1914.

LARRY W. SAMUELSON, Ph.D. Illinois 1978. The marginal product of capital in an aggregate production function.

ROBERT SHERRY, Ph.D. Yale 1979. Petty bourgeois agriculture in the nineteenth century United States.

HASSAN M. SHIRVANI, Ph.D. Harvard 1979. A flow of funds model of the Iranian economy and its implications, 1961-75.

RALPH SHLOMOWITZ, Ph D. Chicago 1979. The transition from slave to freedman: Labor arrangements in southern agriculture, 1865 70.

FEREIDOON SIOSHANSI, Ph.D Purdue 1978. Choice under uncertainty involving matters of life and death. Theory and applications.

DAVID H SMALL, Ph.D Northwestern 1978. A theoretical and empirical examination of the macro-economic rational expectations hypothesis

BARBARA J SPENCER, Ph.D. Carnegie-Mellon 1979. Bureaucratic budgets under asymmetric information: A principal and agent approach

PETER J. STAN, Ph.D Harvard 1979. Welfare economics when transactions are costly

DAVID J. ST CLAIR, Ph D Utah 1979 Entrepreneurship and the American automobile industry.

STEPHEN C SWAIM, Ph.D. Maryland 1979. Theory and policy in the economic thought of Leo Rogin

SUSAN K TOUCHSTONE, Ph D Florida State 1978 A study of the demand for four nonprofit performing arts

RALPH W TRYON, Ph D Massachusetts Institute of Technology 1979 The French franc in the 1920's

THOMAS S ULEN, Ph D Stanford 1978 Cartels and regulation Late nineteenth century railroad collusion and the creation of the Interstate Commerce Commission

STEPHEN E USHER, Ph.D Michigan 1978. Consumer aspiration A dynamic approach.

AHMAD VESSAL, Ph.D Clark 1978. The role of advertising in oligopolistic markets

STEVEN WEBB, Ph.D Chicago 1978 The economic effects of protective tariffs in Germany, 1879 to 1914.

JOHN A. WILLOUGHBY, Ph.D California (Berkeley) 1979 Confrontation or cooperation?: Marxist theories of nation-state competition in the Third World.

JOSEPH S WINRICH, Ph D Kentucky 1978. Environmental externalities and abatement

KIMBERLY D. ZIESCHANG, Ph.D North Carolina (Chapel Hill) 1978. Production scaling laws and the functional structures of production

Economic Growth and Development; including Economic Planning Theory and Policy, Economic Fluctuations and Forecasting

MANMOHAN L. AGARWAL, Ph.D. Massachusetts Institute of Technology 1979. Economic policy in a mineral economy: A case study of Jamaica.

SALEHUDDIN AHMED, Ph.D. McMaster 1978. Rural-urban interactions in economic development: A two-sector model for Bangladesh.

UGUR AKER, Ph.D. Wayne State 1979. Determinants of interprovincial migrations in Turkey between 1960 and 1965.

KADHIM A. AL-EYD, Ph.D. George Washington 1978. Oil revenues, absorptive capacity, and prospects for accelerated growth: A case study of Iraq.

SULAYMAN S. AL-QUDSI, Ph.D. California (Davis) 1979. Growth and distribution in Kuwait: A quantitative approach.

GARY C. ANDERS, Ph.D. Notre Dame 1979. Dependence and underdevelopment: The political economy of Cherokee native Americans.

JACOB K. ATTA, Ph.D. Colorado 1979. A macro-econometric model of Ghana: Simulation experiments in large-scale econometric model building for a developing country.

CESAR M. BARROS, Ph D. Stanford 1979 Money and finance in less-developed countries.

PAUL E. BECKERMAN, Ph.D. Princeton 1979 Essays on the theory of index-linking and its implementation in Brazil.

CHARLES E. BEGLEY, Ph.D. Texas (Austin) 1978. A savings-aid relationship in Latin America.

NATHANIEL B. CABANILLA, Ph.D. George Washington 1979. Money, finance, and growth: An empirical inquiry into the economic development of the Philippines.

DAVID CARR, Ph.D. New York 1979. Impact of foreign direct investment on the economic development of Australia

NORMAN E. CARRUTHERS, Ph.D. California (Berkeley) 1979. Essays on the economics of location choice.

FRANCISCO A. CATALA, Ph.D. Georgetown 1978. The role and behavior of enterprises in the new phase of self-management in Yugoslavia

CHING-HSI CHANG, Ph.D. Ohio State 1978. Female labor force participation in Taiwan.

STEVEN S. CHANG, Ph.D. Georgia State 1978. On a two-sector model of economic growth with money

STEVEN C. CHOW, Ph.D. Boston 1977. Economic growth and income distribution in Hong Kong.

KONSTANTINOS D. CHRISTOU, Ph.D. California (Berkeley) 1979. Greek agricultural development and integration with the European communities.

CARL CLARKE, Ph.D. Fordham 1979. An analysis of the determinants of demand in Barbados.

THOMAS A CONNORS, Ph.D. Michigan 1979 Balance-of-payments adjustment in the West Africa monetary union.

PETER F. CORY, Ph.D. California (Berkeley) 1979. "Technology gaps" in developing countries and the foreign corporation: A comparative study of Yugoslavia with Mexico and Spain.

AERCIO DOS SANTOS CUNHA, Ph.D. Vanderbilt 1978. Rural poverty and agricultural modernization in Brazil.

WILLIAM DARITY, Ph.D. Massachusetts Institute of Technology 1978. Essays in growth, distribution, and development.

HUGH DAWES, Ph.D. New York 1979. The role of public finance in a developing Jamaica.

CARMEN D. DEERE, Ph.D. California (Berkeley) 1979. The development of capitalism in agriculture and the division of labor by sex: A study of the northern Peruvian Sierra.

JULIO A. ECHEVARRIA, Ph.D. Iowa State 1978 A development programming model for the Peruvian agricultural sector.

MUSTAFA S. El-HUNI, Ph.D. Oklahoma State 1978. Economic growth constraints and the role of oil trade in Libya.

RICHARD E. ERICSON, Ph.D California (Berkeley) 1979. Nonmarket resource allocation under uncertainty: Production inputs in a planned economy

ENEFIOK G. ETUK, Ph.D. Michigan State 1979. Microeconomic effects of technological change on smallholder agriculture in northern Nigeria: A linear programming analysis.

JOSEPH FELDER, Ph D California (Los Angeles) 1978 A micro-economic model of fertility in less-developed countries.

YOSHIKAZU FUKASAWA, Ph.D Kansas State 1979. Inflation and business cycles with a special consideration given to price-induced speculative inventories.

HAROLD S. GARDNER, Ph D California (Berkeley) 1978 Efficiency and optimality criteria in Soviet foreign trade planning.

GEORGE C. GEORGIOU, Ph D George Washington 1979 Alternative trade strategies and employment in Cyprus.

DOUGLAS E. GOODMAN, Ph.D Illinois 1978. Oil prices, inflation, and unemployment: A measure of the effect of OPEC policy on *U.S* prices, monetary-fiscal responses, and hence on unemployment rates

ERIC S. GRABER, Ph D. Iowa State 1978. Strategy, policies, and programs for economic growth and social progress in the Dominican Republic: With reference to rural development.

ROBIN H. GRIEVES, Ph.D. Maryland 1979. The demand for consumer durables.

CHRISTIAAN GROOTAERT, Ph.D. Hawaii 1978. Patterns of final demand and income distribution: The case of Japan.

ARUNEE C. GROW, Ph.D. Cincinnati 1978. Macroeconomic model of Thailand: National income model and foreign economic assistance.

AMER A. GUMMED, Ph.D. Oklahoma State 1979. High-level manpower, education, and economic growth in Libya.

ADEEB K. HADDAD, Ph.D. Oklahoma State 1979. An econometric monetary model of the Jordanian economy.

LANA L. B. HALL, Ph.D. California (Berkeley) 1979. Food aid and agricultural development: The case of P.L. 480 wheat in Latin America.

GILLIAN P. HART, Ph.D. Cornell 1978. Labor allocation strategies in rural Japanese households

KARL N. HELLA, Ph.D Washington (St. Louis) 1978 Surplus labor and the production of nonmarket goods in rural Vietnam.

JOHN K. HILL, Ph.D. Rice 1979. Capital markets and household saving: Empirical tests and simulations for urban Columbia.

CHUL HONG, Ph.D. Pennsylvania 1979. A study of physical capital investment in Korea.

JOHN C. ICKIS, D.B.A. Harvard 1979. Strategy and structure in rural development.

MAARTEN IMMINK, Ph.D. Hawaii 1978. Human energy supplementation and worker productivity. A case study of sugarcane cutters in Guatemala

ELIJAH M. JAMES, Ph D California Western 1978 Regional economic integration and the development of the mini economy.

ROBERT E. JORDAN, Ph D. Texas (Austin) 1979. An evaluation of public housing policy in Peru.

MUSTAPHA KARA, Ph.D. Pittsburgh 1978. Fiscal policy and economic development· A case study, Algeria

SELIM G. KARATAS, Ph.D. Oklahoma 1978 The demand for money in a fragmented economy. The case of Turkey.

MUHAMMAD F. KHAN, Ph D Boston 1979 A study of the causes of fluctuations of real wages in the labor surplus economy of Pakistan.

SULEIMAN I. KIGGUNDU, Ph.D. Boston 1979 The rationalization of industrial economic planning and development in the East African common market: A game theoretical approach.

IN JUNE KIM, Ph.D Harvard 1979. The impact of increase in input prices on the growth rate of output and inflation.

SI DAM KIM, Ph.D. State University of New York (Binghamton) 1978 Interest rate reform in a financially underdeveloped economy: The case of Korea

ULRICH LACHLER, Ph.D. Columbia 1979. Essays in the theory of political business cycles.

YEW-WAH LAI, Ph.D. Hawaii 1978. Sources of growth and employment creation in the Malaysian manufacturing industries.

ANDRE P. LARA-RESENDE, Ph.D. Massachusetts Institute of Technology 1979. Inflation, growth, and oligopolistic pricing in a semi-industrialized economy: The case of Brazil.

PHILLIP G. LEBEL, Ph.D. Boston 1977. A linear planning model for education with application to Senegal.

VICTOR C. LEVY, Ph.D. Chicago 1978. Aspects of factor productivity and economic efficiency in a socialist developing economy.

DAVID M. LILIEN, Ph.D. Massachusetts Institute of Technology 1977 The cyclical pattern of temporary layoffs in the United States.

LINDA Y. LIM, Ph.D. Michigan 1978. Multinational firms and manufacturing for export in less-developed countries: The case of the electronics industry in Malaysia and Singapore.

YANK TAEK LIM, Ph.D. Georgia State 1978. Optimal economic growth and social security.

DAVID L. LINDAUER, Ph.D. Harvard 1979. Labor market behavior in developing nations.

CARLOS A. LONGO, Ph.D. Rice 1979. Border tax adjustments and the allocation of tax revenues: The case of the state value-added tax in Brazil..

NORA C. LUSTIG, Ph.D. California (Berkeley) 1979. Distribution of income, structure of consumption, and economic growth: The case of Mexico

JOHN H. MCDERMOTT, Ph.D. Brown 1979. Optimal resource depletion and capital accumulation in developing economies.

CHRISTIAN M. MADU, Ph.D Colorado 1979. An econometric approach to forecasting the recurrent revenue of a developing country· The Nigerian case.

KOSEI MANAWAKALAO, Ph.D Iowa State 1978. A crop and bovine model of Thailand: A spatial analysis.

AMIT K MITRA, Ph.D Duke 1978 An intertemporal cross-country analysis of the impact of economic and demographic factors on government expenditure share

SHAHRIAR MOSTASHARI, Ph.D North Carolina State 1978 Analytical survey of the growth of the Iranian economy. A macro-econometric model.

SALEM M. MOUSTAFA, Ph D. Southern Methodist 1979. A model of the Libyan economy.

BENNO J NDULU, Ph D. Northwestern 1979. The role of transportation in agricultural production: The case of Tanzania

SAMUEL M. NGOLA, Ph.D New Mexico 1979. Central banking in East Africa.

NICHOLAS R. NOBLE, Ph D Cincinnati 1978 The formation of inflationary expectations: The role of government policy

JOHN L. NUTTALL III, Ph.D. Illinois 1978 Estimating the gross value of industrial and agricultural output of the People's Republic of China.

CHRISTIAN ODOFIN, Ph.D. American 1979. The impacts of the multinational oil corporations on Nigeria's economic growth: Theoretical and empirical explorations.

CHARLES P. OMAN, Ph.D. California (Berkeley) 1978. Rural-urban migration and economic underdevelopment in Peru.

DANIEL E. O'NEILL, Ph.D. Massachusetts Institute of Technology 1977 Directly estimated multipliers of monetary and fiscal policy.

THEODORE PANAYOTOU, Ph.D. British Columbia 1978. The contribution of nonrenewable natural resources to economic development: The case of copper in Zambia.

TIRSO B. PARIS, Ph.D. Michigan State 1979. Systems analysis and simulation of up-land river based cropping systems in the Philippines.

AE SIL KIM PARK, Ph.D. Hawaii 1978. Technological progress, diffusion of technology, and international differences in labor productivity.

SEUNG RHEE, Ph.D. Michigan State 1978. Multi-products production relations in manufacturing plants: An explanatory study on six selective manufacturing activities in Korea.

ALEXANDER RIDDLE, JR., Ph.D. Pittsburgh 1979. A test of Jorgenson's alternative theories of the development of a dual economy.

NICHOLAS H. RIEGG, Ph.D. Connecticut 1979. The role of fiscal and monetary policies in Taiwan's economic development.

ROBERT D. ROE, Ph.D. Duke 1978. A theoretical and empirical examination of public sector scale economies.

GEORGE M RUIGU, Ph.D. Michigan State 1978. An economic analysis of the Kenya milk subsystem.

PREECHA SAKARINDR, Ph.D. Iowa State 1979. An econometric study of Thai rubber industry and the world rubber markets.

SYED A. SAMAD, Ph.D. Boston 1979 Intertemporal movements of real wages. A case study of Bangladesh, India, and Indonesia.

NABIL SAMMAN, Ph.D Cincinnati 1978 Cost-benefit assessment of the Euphrates dam project

ALFRED SANT, D.B.A. Harvard 1979. Problems of government organization and management in the state enterprises of developing countries The Maltese case

CAYETANO S. SARMAGO, Ph.D Michigan State 1978 Analysis of the rural banking system in the Philippines.

RAPEEPUN SEKTHEERA, Ph D Michigan State 1979. The allocation of family resources to farm and nonfarm activities in a village in northern Thailand.

ASIM SEN, Ph.D. Rutgers 1979 The role of technological change in economic development: The lessons of Japan for presently developing countries.

JOSÉ J. SIDAOUI, Ph.D. George Washington 1979. Factor-market distortions in a multisectoral model: A case study of Mexico.

BALBIR S. SIHAG, Ph.D. Massachusetts Institute of Technology 1978. Economics of foodgrain reserves for India.

DEVADASON D. D. SOLOMON, Ph.D Hawaii 1978. Population policy and employment strategy for the development of West Malaysia.

JARILAOS STAVROU, Ph.D. Iowa State 1978. A macroeconometric model for the Panamanian economy, 1950–72.

CHARLES STEINDEL, Ph.D. Massachusetts Institute of Technology 1977. Personal consumption, property income, and corporate saving.

NEAL C. STOLLEMAN, Ph.D City (New York) 1978. An economic growth model with endogenous labor supply and savings functions.

LEONORE S. TAGA, Ph.D. California (Berkeley) 1979. Externalities in the Soviet economy. Forestry problems and policies.

ABDULLAH M. TAHER, Ph D Illinois 1978. External borrowing and economic growth in Jordan during the period 1955-75.

SHIN-ICHI TAKEKUMA, Ph.D. Rochester 1979. Mathematical theory of optimal economic growth.

SAMUEL TORRESROMAN, Ph.D. Michigan State 1979. Structural transformation in the economy of Puerto Rico from 1940 to 1972 An input-output analysis

THOMAS J. TREBAT, Ph.D. Vanderbilt 1978. An evaluation of the economic performance of public enterprises in Brazil

SHUJIRO URATA, Ph D. Stanford 1978 Effects of protection on resource allocation and economic welfare—The Japanese case

LUIS VALDIVIESO MONTANO, Ph D. Boston 1979. The distributive effect of alternative policies to increase the use of existing industrial capacity.

A. VAN POECK, Ph.D Katholieke Unversitet Leuven 1979 World price trends and price and wage developments in Belgium

MINOR VARGAS, Ph.D McMaster 1978 Financial interdependence and economic policy in Costa Rica.

VICTORIANO M VARGAS, Ph.D. Michigan State 1978 Tax expenditures in Panama, 1975: An empirical study

AMY L. WALTON, Ph.D Princeton 1979 Energy prices, investment, and economic growth in the Middle Atlantic Region

DENNIS L. WARNER, Ph.D Princeton 1979 An econometric model of the world wheat economy

MICHAEL A WATTLEWORTH, Ph.D California (Berkeley) 1979. Agricultural development through technological change and diversification: An economic analysis of the rubber estate sector of peninsular Mayalsia, 1950's to 1970's.

ROBERT J WHITACRE, Ph.D. California (Santa Barbara) 1978. Simulation and control in macro-economic growth theory.

CHRISTINE WONG, Ph.D. California (Berkeley) 1979 Rural industries in the People's Republic of China.

DARREL A. YOUNG, Ph.D. Texas (Austin) 1979. Men vs. machines: A study of choice of technology in Andean metalworking firms.

ALY ZEINELDIN, Ph.D. New York 1979. The future of the Egyptian economy and the choice between alternative scenarios: An input-output advent

Economic Statistics; including Econometric Methods, Economic and Social Accounting

SULTAN AHMAD, Ph.D. Pennsylvania 1978. Shortcut methods of international comparison of real product and purchasing power of currencies.

NARIMAN BEHRAVESH, Ph.D. Pennsylvania 1978. Spectral estimation of dynamic econometric models with serially correlated errors.

STANLEY K. BERRY, Ph.D. Vanderbilt 1979. Small-sample properties of four pseudo-disturbance generators.

SALAH BRIK-EL-HANNACHI, Ph.D. Columbia 1979. A decision-oriented formalization of power.

MICHAEL F. CHARETTE, Ph.D. Western Ontario 1978. Sampling properties of the Theil mixed estimator

STEPHEN R. COSSLETT, Ph.D. California (Berkeley) 1979. Estimating choice probabilities from choice-based samples: Efficient estimators and their application to travel demand forecasting

ALAIN J.-M. D. B. DE FONTENAY, Ph.D. Vanderbilt 1979. A contribution to the foundation of seasonal analysis

JEAN-MARIE DUFOUR, Ph.D. Chicago 1979. Methods for specification error analysis with macro-economic applications.

DANIEL EGY, Ph.D. State University of New York (Albany) 1979. Nonspherical disturbances and functional form problems in econometric models.

ROMAN FRYDMAN, Ph.D. Columbia 1978. Non-linear estimation problems

SYAMAL K. GHOSH, Ph.D. Iowa State 1979. Designing a pilot set of social accounts based on the Fox-Van Moeseke total income approach.

JOSEPH S. GRAVES, Ph.D. Carnegie-Mellon 1979. A class of set partitioning problems with additional structure: Some applications and approaches to solution

DINH T. HINH, Ph.D. Pittsburgh 1978. Autocorrelation in simultaneous equations models.

CHIH-HO HUANG, Ph.D. Illinois 1978. An econometric study of the consumers' demand for passenger cars in the United States.

ROSELYNE JOYEUX, Ph.D. California (San Diego) 1979. Harmonizable processes in economics

R. KRISHNA KADIYALA, Ph.D. Western Ontario 1978. The finite sample properties of ridge estimators.

SEIICHI KAWASAKI, Ph.D. Northwestern 1979. Application of log-linear probability models in econometrics.

MOHAMMED S. KHALED, Ph.D. British Columbia 1978. Productivity analysis and functional specification: A parametric approach

PHILLIP S. KOTT, Ph.D. Brown 1979. The study and estimation of particular technologies consistent under hypothesized technical progress functions.

WILLIAM S. KRASKER, Ph.D. Massachusetts Institute of Technology 1978. Applications of robust estimation to econometric problems.

MILES C. LEAHEY, Ph.D. Pennsylvania 1978. Economic impact of pollution control on the paper industry: An econometric analysis.

HERMAN B. LEONARD, Ph.D. Harvard 1979. Inference in partially identified models.

CLARENCE H. MARTIN, Ph.D. Carnegie-Mellon 1979. Computational experiments with zero-one programming algorithms and heuristics.

RICHARD A. MEESE, Ph.D. Wisconsin (Madison) 1978. Distributed lag order determination with an application to the multiperiod theory of the firm.

PETER J. MEYER, Ph.D. California (Berkeley) 1979. The reproduction of the distribution of income.

VINAY PANDIT, Ph.D. Columbia 1978. Data-breeder: Estimation of microdata in an incomplete data base.

DAVID M. PRESCOTT, Ph.D. Queen's (Kingston) 1978. Weighted ridge regression under the objectives of estimation and conditional prediction.

CHARLES S. ROEHRIG, Ph.D. Michigan 1978. Estimation of *M*-equation linear models subject to a linear constraint on the endogenous variables

SAMUEL SCHWARTZ, Ph.D. Columbia 1978. Estimation of a truly complete system of demand equations: An application to a gambling market.

SUSAN A. SIMMONS, Ph.D. Mississippi 1976. Estimation of a vintage model of production for the *U.S.* pharmaceutical preparations industry.

JEFFREY K. SPEAKES, Ph.D. California (Berkeley) 1979. Identification and efficient estimation in disaggregated models with special covariance structure.

THOMAS M. STOKER, Ph.D. Harvard 1979. Aggregation over individuals and demand analysis: A theoretical and applied study.

AKIRA TOMIOKA, Ph.D. Stanford 1978. A study on persistent disequilibrium

MICHAEL WARD, Ph.D. Chicago 1978. Optimal production and inventory decisions: An analysis of firm and industry behavior

MARK D. WESTBROOK, Ph.D. Ohio State 1978. An experimental investigation of the properties of simultaneous equations estimators.

G. THOMAS WOODWARD, Ph.D. Brown 1979. An aggregate "world" production function: Theory and estimation.

BONG JOON YOON, Ph.D. Illinois 1978. The duration of unemployment: A search information approach.

CHANG-HO YOON, Ph.D. Stanford 1978. Rational expectations equilibrium, cost of information, and welfare.

Monetary and Fiscal Theory, Policy, and Institutions

ANDREW B. ABEL, Ph.D. Massachusetts Institute of Technology 1978. Investment and the value of capital.

RICHARD K. ABRAMS, Ph.D. North Carolina (Chapel Hill) 1978. Monetary policy reaction functions using a random coefficient regression model.

AHMAD A. M. ABU-SHAIKHA, Ph.D George Washington 1979. A theoretical analysis of the impact of fiscal and monetary policies on economic activity in closed and open economies.

KATHLEEN E. ADAMS, Ph.D. Colorado 1978. Analysis of fiscal response to school districts to incentives of the state aid formula in Colorado.

LORETO P. ALONZI, Ph.D. Iowa 1979. Dynamic aspects of uncertainty, portfolio choice, and financial intermediation.

MICHAEL R. ASAY, Ph.D. Southern California 1978. The theory of rational mortgage pricing: An application of contingent claim.

SALLY U. ASEFA, Ph.D. Iowa State 1978. The determinants of cross-sectional variation in city borrowing.

JOSEPH C. ATKINS, Ph.D. Pennsylvania State 1978. A conceptual and empirical analysis of commercial bank portfolio management.

ZETI A. AZIZ, Ph.D. Pennsylvania 1978 Commercial bank portfolio behavior in an open developing economy: The Malaysian case.

MARIO BALDASSARI, Ph.D Massachusetts Institute of Technology 1978. Government investment, inflation, and growth in a mixed economy: Theoretical aspects and empirical evidence of the experience of Italian government corporation investments.

ANTONIO BARBOSA, Ph.D. Virginia Polytechnic Institute 1978. The constitutional approach to the fiscal process. An inquiry into some logical foundations

LUIS M. C. P. BELEZA, Ph.D. Massachusetts Institute of Technology 1979 Credit, wages, and prices in a small, open economy: Portugal

KENNETH M BERNAUER, Ph D Illinois 1978 Dynamic modelling of expectations in foreign exchange markets.

DONALD R. BONNEY, Ph.D. Iowa 1978. Capitalization of intrajurisdictional differences in the property tax

JEREMY BULOW, Ph.D. Massachusetts Institute of Technology 1979. Essays in macroeconomics and finance.

JOHN B. CARLSON, Ph.D. Michigan State 1978. The bank loan decision under the expected utility hypothesis

ALBERTA H. CHARNEY, Ph.D Illinois 1979. Local taxes, public services, and intrametropolitan industrial complex.

RONALD L. CHASTAIN, Ph.D. Washington State 1979. An empirical evaluation of Oregon's property tax relief program.

SOON BENG CHEW, Ph.D. Western Ontario 1979. The effects of incomes policy on U K. wage inflation at the industrial level, 1948–75.

HENRY A. COLEMAN, Ph.D. Princeton 1979 Government by special districts.

PATRICK S. COTTER, Ph.D. Wisconsin (Milwaukee) 1978. Implicit macro-economic goals: An application of the inverse problem of optimal control.

JOSEPH C. COX, Ph.D. McMaster 1979. The interregional impact of federal grants to provincial governments.

TIMOTHY J. CURRY, Ph.D. George Washington 1979. Bank holding companies as financial conglomerates and their effect upon bank behavior.

CHARLES J. DALE, Ph.D. Georgia State 1978. An empirical comparison of the predictive ability of monetarist and structural macro models.

JOHN J. D'ANDREA, Ph.D Princeton 1978. The mortgage, housing, and deposit sectors of the Canadian economy.

LESLEY D. DANIELS, Ph.D. Stanford 1979. The role of local government in economies with local public goods.

ENID S. DENNIS, Ph D. State University of New York (Binghamton) 1979. A neoclassical analysis of the demand for real cash balances by firms.

ROBERT J. EDGAR, Ph.D. Ohio State 1978. Regional lags in deposit adjustment.

AHMED S. EL-SHARIF, Ph.D Indiana 1979. A structural model of the monetary sector of Libya.

FRANK A. ELSTON, Ph.D. Virginia 1979. Nonmember banks, reserves, and monetary control

RICHARD E FEINBERG, Ph D. Stanford 1978. The Export-Import Bank in the U.S economy

RADU A FILIMON, Ph D. Carnegie-Mellon 1979 Aggregation and the allocation of private and public goods.

DENNIS J. FIXLER, Ph D Purdue 1978 The incentive to invest.

CLAUDE A. FONGEMIE, Ph D Connecticut 1979 A dynamic model of a commercial bank

EDWARD A FRIEDMAN, Ph D Yale 1978 Inflation and international capital flows. A portfolio approach.

A. F FRIES, JR., Ph.D. Lehigh 1978. Inflation and the progressivity of the U.S. tax structure, 1959–72.

FREDERICK T FURLONG, Ph.D California (Los Angeles) 1978 Banking structure and the theory of economic regulation.

BONNIE GARRETT, Ph D. George Washington 1979 Bank and nonbank involvement in the immediately available funds market.

CYNTHIA F GENSHEIMER, Ph D. California (Los Angeles) 1979. The political economy of federal project grant distribution.

WILLIAM D GERDES, Ph D Nebraska (Lincoln) 1978 The federal funds market. A study of the participation by country banks, 1965–75.

ROBERT P. GOODMAN, Ph D. Washington (St. Louis) 1978. Interstate and intertemporal variations in the provision of municipal public services.

CHARLES M. GRAY, Ph.D. Washington (St Louis) 1978. Liability management in a commercial bank model.

RIK W HAFER, Ph.D. Virginia Polytechnic Institute 1979. An empirical comparison of autoregression and rational models of price expectations.

JAMES D. HAMMOND, Ph.D. Virginia 1979. The household demand for liquid assets.

FRANCIS J. HARMAN, Ph.D. McMaster 1977. An analysis of investment incentive policies in Canada.

ROBERT B. HARRIS, Ph.D. Ohio State 1979 Policy options in the taxation of commercial banks with special reference to Ohio.

LOYD J HELMS, Ph.D. Massachusetts Institute of Technology 1979. Optimal taxation and social security in a stochastic economy.

RAYMOND D. HILL, Ph.D. Massachusetts Institute of Technology 1978 Capital market equilibrium and the regulation of property-liability insurance

YAN KI HO, Ph.D. Wisconsin (Madison) 1979. Commercial banking and regional growth: The Wisconsin case.

DAI-SHENG HONG, Ph.D. Minnesota 1978 A method of measuring the finance charge on consumer installment loans.

SRINIVASA INGUVA, Ph.D Georgia State 1978 The demand for money and inflationary process An econometric model.

CHRISTOPHER M. JAMES, Ph D. Michigan 1978 A theory of bank portfolio allocations under uncertainty and the implications for bank regulations

ROBERT G JAMES, Ph D Oregon 1978 The permanent-income hypothesis and other approaches to aggregate consumption functions.

SOO MYUNG JANG, Ph D Cincinnati 1978 The impact of legal reserve requirement changes on excess reserves

GIROL KARACAOGLU, Ph.D. Hawaii 1978 A monetarist model with an aggregate transaction variable and endogenous money supply

HOWARD KEEN, JR., Ph D Bryn Mawr 1979 Household demand for demand deposits. Do nonpecuniary payments make a difference?

PETER M. KERR, Ph D Kansas 1978 A guide for monetary policy

LLOYD R KEYSER, Ph.D Temple 1978 An alternative approach to the corporate demand for money An empirical investigation

MOHAMMAD KHAN, Ph D Oklahoma State 1979 Output and price effects of the automatic fiscal stabilizers during the period 1966-75

RICHARD S. KRASHEVSKI, Ph.D. Georgetown 1978 Nonfinancial corporate business liquid asset holdings

ANDREA L. KUSKO, Ph D. Stanford 1979 Voting, bargaining, and factor income distribution.

GARY A. LATANICH, Ph D Nebraska (Lincoln) 1978 Federal water and sewage grants. The implied urban policy.

L. DOUGLAS LEE, Ph.D. George Washington 1978 The Congressional Budget Act of 1974: The impact on fiscal policy.

YOUNG SAE LEE, Ph.D. Pennsylvania 1978. A disequilibrium model of the commercial loan market under uncertainty: The case of Korea.

CAROL LEISENRING, Ph.D. Michigan State 1978. The structure of reserve requirements and monetary control.

MING-YIH LIANG, Ph.D. Harvard 1979. The involuntary demand for money and the case of the missing money.

JAMES T. LINDLEY, Ph.D. Georgia 1977. An analysis of the Federal Advisory Council of the Federal Reserve System, 1914-38.

JOSEPH R. LIRO, Ph.D Syracuse 1979. A model of state government fiscal behavior: The case of New York.

THOMAS J. MCCOOL, Ph.D. Columbia 1979. Government expenditures and subsidies financed by factor taxes and tariffs in a small, open economy.

WILLIAM D. MCMILLIN, Ph.D. Louisiana State 1979. A theoretical and empirical analysis of the relationship between fiscal policy and the money supply.

ROBERT J MARJOLIN, Ph.D. Harvard 1979. Productive uncertainty, efficiency, and decentralization: A money-equity market organization.

ROBERT D MARTIN, Ph.D. Southern Methodist 1978. The household demand for money.

RANDALL C MERRIS, Ph.D Kentucky 1979. Monetary policy, bank regulation, and banking profits.

ROSS D MILBOURNE, Ph.D California (Berkeley) 1978. Essays on macro-economic and monetary theory.

MICHAEL A MOGAVERO, Ph D. Connecticut 1979. Effect of tax expenditures on individual income tax structure An examination of vertical and horizontal costs

VINCENT G MUNLEY, Ph.D State University of New York (Binghamton) 1979 An economic analysis of sequential referenda for the finance of local public schools.

THOMAS B NEVILLE, Ph D. Stanford 1979. Inflation financed steady-state tax effects

ROBERTO E. NEWELL-GARCIA, Ph.D. Texas (Austin) 1979 Financial deepening and financial narrowing: The case of Mexico, 1960-76.

EDWARD OFFENBACHER, Ph.D. Chicago 1979. Measuring the moneyness of monetary assets and the demand for money

JANUSZ ORDOVER, Ph D. Columbia 1978. Three essays on economic theory

ANTHONY M PAGANO, Ph.D. Pennsylvania State 1978. Measuring the output of state and local units of government.

THOMAS J. PARLIMENT, Ph.D. State University of New York (Binghamton) 1979. An analysis of fiscal consolidation of school districts.

ROBERT W. PARRY, JR., Ph.D. Lehigh 1978. The effects of lease capitalization on the debt ratios of selected Pennsylvania cities.

JAMES A. PIHERA, Ph.D. Georgia State 1979. A reaction function for the Federal Reserve System, 1974 to the present.

SUSAN RANNEY, Ph.D. Wisconsin (Madison) 1978. Capital market imperfections in a model of house purchase decisions.

CHARLES S. SALLEY, Ph.D. Georgia State 1978. Spatial differentiation and price: A test of delineated banking markets.

ROBERT W. SAMOHYL, Ph.D. Rice 1979. The question of money substitution in Brazil.

JIA-DONG SHEA, Ph.D Stanford 1978. The policy implications of the government budget constraint in an optimal accumulation model.

NEIL SHEFLIN, Ph.D. Rutgers 1979. Forecasting macro-financial behavior: A bank reserves flow-of-funds approach.

MILAN P. SIGETICH, Ph.D Stanford 1979. The financial effects of deregulation on California savings and loan associations.

PAUL A SPINDT, Ph.D. California (Santa Barbara) 1977. Deposit expectations, forecasting, and individual bank balance sheet dynamics

RICHARD STARTZ, Ph D. Massachusetts Institute of Technology 1978. Interest-bearing demand deposits

GORDON L STREEB, Ph D. Minnesota 1978 The distribution of local government services: A case study of Minneapolis.

JACK P. SUYDERHOUD, Ph D Purdue 1978 The motivation for and voter response to state tax expenditure limitations—A descriptive, theoretical, and empirical analysis.

BADROSSADAT TAHADJOD, Ph.D Clark 1978 Are households rational or ultrarational? A theoretical and empirical inquiry.

JOSEPHINE K. TAN, Ph.D Alabama 1979 The determination of interest rates: An evaluation of the specification of alternative model structures

MOHAMMED A. TAREEN, Ph D. Pennsylvania 1978 A theory of money supply and its macro-econometric implementation. Description, stabilization policy

VEFA TARHAN, Ph.D. California (Santa Barbara) 1977 A micro-economic analysis of commercial bank portfolio behavior.

JOSEPH M TERRY, Ph.D. California (Santa Barbara) 1979. Measuring and comparing the adequacy of a sales tax and a personal income tax using California data.

ALFREDO VERGARA, Ph.D. Carleton 1979. Money in economic development: The Ecuadorian case.

PETER W. VITALIANO, Ph.D. Wisconsin (Madison) 1979 An empirical analysis of the cost of term borrowing for agricultural marketing cooperatives and proprietary agricultural marketing businesses

BRUCE G. WEBB, Ph.D. Clark 1979. Alternative mortgage instruments and borrower risk.

ROBERT I. WEBB, Ph.D. Chicago 1978 Impact of open market operations on Treasury Bill yields.

JOHN D. WILSON, Ph.D Massachusetts Institute of Technology 1979. Aspects of optimal income taxation and public employment.

JAMES P. WINDER, Ph.D. Purdue 1978. The determinants of commercial bank assets, liabilities, and income.

SCOTT WINNINGHAM, Ph.D. Wisconsin (Madison) 1978. Liability management banking.

HIROSHI YOSHIKAWA, Ph D. Yale 1978. An essay on macroeconomics and monetary policy

NAOYUKI YOSHINO, Ph.D. Johns Hopkins 1979. Optimal choice of monetary and fiscal policy under uncertainty

TEH-PEI YU, Ph D Hawaii 1978 Balance of payments and monetary control in the Republic of China

EMAD ZIKRY, Ph D. Kansas 1979. A monetary growth model with unemployment and inflation

DENNIS P ZOCCO, Ph D Lehigh 1978 The corporate income tax. Its evolution and justification.

BENJAMIN ZYCHER, Ph D California (Los Angeles) 1979 Substitution in public spending The budgetary incidence of Canadian national health insurance

International Economics

CHARLES M AHO, Ph D Massachusetts Institute of Technology 1978 The overvaluation of the dollar and *U.S.* foreign direct investments

HIROYA AKIBA, Ph D State University of New York (Binghamton) 1978 Nontraded goods and the gains from trade under price uncertainty

DEBORAH L ALLEN, Ph D Duke 1978 Tariff games

PEDRO A ASPE, Ph D Massachusetts Institute of Technology 1978 Essays on the international transmission mechanism The Mexican case.

MICHAEL BAZDARICH, Ph D Chicago 1978 Economic development, optimal growth, and stages in the balance of payments

JULIAN BERENGAUT, Ph D. Wisconsin (Madison) 1978 International commodity arbitrage and the relationships between foreign and domestic prices in Canada and the United States.

ADMASSU BEZABEH, Ph.D California (Berkeley) 1978 Balance-of-payments constraint and economic development of Ethiopia, 1960–74

ERIC W BOND, Ph D. Rochester 1979. Taxation and multinational firm behavior: Some evidence from Puerto Rico

RICHARD BOOKSTABER, Ph D. Massachusetts Institute of Technology 1977. The international role of financial markets.

MARIANNE BOWES, Ph.D. Illinois 1978. Spatial competition and the theory of international trade.

JAMES A. BRANDER, Ph.D. Stanford 1979. The division of labor and intra-industry trade.

ANDRE C. BURGSTALLER, Ph.D. Columbia 1979. Flexible

exchange rates, quasi-rational expectations, and variations in the rate of unemployment.

JORGE CAMBIASO-ROPERT, Ph.D. Wisconsin (Madison) 1978. An empirical analysis of money and the balance of payments in Latin America.

LEONARDO A. O. CARDEMIL, Ph.D. Vanderbilt 1978. A disequilibrium model for the short-run macroeconomic effects of exchange rate variations

ANTHONY CASSESE, Ph.D. City (New York) 1979 Money, the dynamics of inflation, and the balance of payments in Latin America, 1947–76.

DON-SE CHA, Ph.D Vanderbilt 1979. International short-term capital flows, exchange rate expectations, and independence of monetary policy.

NIKHIL A. CHANDAVARKAR, Ph D Georgetown 1978 The balance of payments and inflation in Brazil: A quantitative study.

LIANG CHANG, Ph D State University of New York (Albany) 1978. Effect of official transfer payments in a two-country portfolio model.

RAYMOND K CHIEN, Ph D. Pennsylvania 1978. Economic integration among less-developed countries: Theory and ASEAN practice.

DAVID T. COE, Ph D. Michigan 1978 Impact effects of policy in a two-country portfolio-balance model.

CARLOS N. COEHLO, Ph.D. Missouri (Columbia) 1978. An economic analysis of the Brazilian value-added tax.

I. LILLA CONNIDIS, Ph.D Queen's (Kingston) 1979 A study of the auto pact and its consequences.

ISMAIL DAIS, Ph.D. City (New York) 1979 The effect of trade liberalization on national specialization An empirical study of changes in trade patterns and industrial structure in West Germany, France, and Netherlands

WILLIAM A. DELLALFAR, Ph D Massachusetts Institute of Technology 1978 Sweden's choice of a currency regime.

SLOBODAN DJAJIC, Ph.D. Columbia 1979 Flexible exchange rates in a world of perfect capital mobility. A rational expectations approach.

SHABTAI DONNENFELD, Ph.D. Indiana 1979. International trade under uncertainty.

JULIO H. DURAN, Ph.D. Minnesota 1978. Balance of payments in Spain: An econometric study.

JOHN C DUTTON, JR., Ph.D. Duke 1978. Effective protection, taxes on foreign investments, and the operation of the gold standard.

ARON DYCHTER, Ph.D. George Washington 1979. Flexible exchange rates and the less-developed countries' direction of trade.

CHARLES A. ENOCH, Ph.D. Princeton 1979. Optimal incentives in manufacturing for export.

YOON-DAE EUH, Ph.D. Michigan 1978. The credit worthiness of *LDC*s: The perception of commercial banks.

GREGORY B. FAGER, Ph.D. Georgetown 1978. The recent experience with floating exchange rates: A statistical examination of daily exchange-rate movements.

MAHNAZ FAHIM-NADER, Ph.D. Maryland 1978. Commercial policies and resource allocation: Case study of Iran.

FARHAD FOZOUNI, Ph.D. California (Santa Barbara) 1977. Uncertainty and insurance in the pure theory of international trade.

JEFFREY A. FRANKEL, Ph.D. Massachusetts Institute of Technology 1978. Exchange rates since floating: Theory and evidence.

HAROLD O FRIED, Ph.D. North Carolina (Chapel Hill) 1979 Essays on content protection under world monopoly

HENRI GHESQUIERE, Ph.D. Yale 1979 Causes and implications of the growth of the Euro-dollar market.

FRANCESCO GIAVAZZI, Ph.D. Massachusetts Institute of Technology 1978. Trade flows, relative prices, and the exchange rate in the short run.

REUVEN GLICK, Ph.D. Princeton 1979. Multinational firms and exchange rates: Intertemporal optimization under uncertainty.

ROY T. GOBIN, Ph.D. Illinois 1978. An analysis of indirect taxation within the Caribbean Common Market.

LAURIE GOODMAN, Ph.D Stanford 1978. Equilibrium of bank foreign exchange trading.

ROBERT GREENFIELD, Ph.D Rutgers 1979. Some international evidence on the demand for competitively remunerated quality-adjusted money balances.

SOMPORN HANPONGPANDH, Ph D Iowa State 1978 The ASEAN rice security reserves: A simulation approach

JAMES E HODDER, Ph D Stanford 1978. The hedging of exposure to exchange-rate movements.

MOTOSHIGE ITOH, Ph.D. Rochester 1979. A theory of imperfect competition in international trade and investment

ARVIND K. JAIN, Ph.D. Michigan 1979. International integration of commodity markets A test of arbitrage under changing exchange rates.

JAMES JOHANNES, Ph.D Wisconsin (Madison) 1978. Worldwide inflation and *U.S.* monetary policy: Tests of the structural specification and implications of a monetary balance-of-payments model.

WILLIAM H. KAEMPFER, Ph.D. Duke 1979. The effects of protection and transport costs on the composition of trade.

GEORGE KANATAS, Ph.D. John Hopkins 1979. Essays on exchange-rate dynamics.

MENACHEM KATZ, Ph.D Columbia 1979. External indebtedness and default in a developing economy.

MASAHIRO KAWAI, Ph.D. Stanford 1978. International investment in a world of floating exchange rates: A micro-economic perspective.

MOHAMMAD Z. KHAN, Ph.D. Johns Hopkins 1979. The system of export incentives in the manufacturing sector of Pakistan.

KI HO KIM, Ph.D. Rutgers 1979. A study of the factors affecting Japan's trade balance, 1974–76.

HARALAMBOS S.-M. KOSTAKOPOLOUS, Ph.D. Columbia 1979. An econometric model of the Greek economy with application to measuring the trade effects of the customs union of Greece with the European Economic Community.

PAUL R. KRUGMAN, Ph.D. Massachusetts Institute of Technology 1977. Essays on flexible exchange rates.

ROBERT Z. LAWRENCE, Ph.D. Yale 1978. The measurement and causes of the synchronization of the international business cycle.

JONG WON LEE, Ph.D. Indiana 1978. The formation of international cartels.

PIBOON LIMPRAPAT, Ph.D. Illinois 1978. The major factors affecting instability of export proceeds in Thailand, 1960–76.

ANSELM L. LONDON, Ph.D. Queen's (Kingston) 1978. The informational efficiency of the Canada-*U.S.* foreign exchange market.

MONTAGUE J. LORD, Ph.D. Texas (Austin) 1979. The comparative national performance in instability and growth of Latin America's primary commodity exports.

JOHN J. MCAULEY, Ph.D. Fordham 1979. A study of the impact of general price level changes (inflation rates) on exports transmitted through the factor-price ratio in four developed economies.

JULIET MAK, Ph.D. Wisconsin (Madison) 1978. Alternative approaches to industrialization. A comparative analysis of the Philippines, South Korea, and Taiwan

PURAYIDA I. MATHEW, Ph.D. Connecticut 1979. The Soviet Union's economic relations with India during 1951–76: A study in political economy.

STEPHEN MEYER, Ph.D. Yale 1979. Macro-economic policy in large, open economies with floating exchange rates.

MARIE-CHRISTINE MEYER-ADAM, Ph.D. Johns Hopkins 1979. Capital flows and government intervention in the foreign exchange market: The experience of the Belgo-Luxemburg economic union.

ROSEMARY E. MINYARD, Ph.D. Princeton 1979. Industry and host-country incentives for foreign direct investment: Multinational corporations.

ERNESTO MIRANDA, Ph.D. New Hampshire 1979. A model for the allocation of foreign aid.

AKBAR NAFARI, Ph.D. Indiana 1978 *U.S.* multinational corporations in Mexican manufacturing: A study of development and balance-of-payments impacts.

GEORGE N. NASKARIS, Ph.D. City (New York) 1979. Monetary aspects of the balance-of-payments theory: The Greek experience, 1957–76.

MOHSEN NOORBAKHSH, Ph.D. California (Davis) 1979. Dynamic consideration of trade agreements: Case study—Iran.

ARVIND PANAGARIYA, Ph.D. Princeton 1978. The pure theory of international trade and variable returns to scale.

LARS PEDERSON, Ph.D. Columbia 1979. The optimal movement of factors in a two-sector model of production.

JOHN J. PIDERIT, Ph.D. Princeton 1979. Stochastic commodity stocks in international trade.

PRADUMNA B. RANA, Ph.D. Vanderbilt 1979. The impact of generalized floating on trade flows and reserve needs: Selected Asian developing countries.

ROMUALDO A. ROLDAN, Ph.D. Pennsylvania 1978. The technology of multinational corporations. A statistical analysis of their factor intensities.

NASSER A. H. SAIDI, Ph.D. Rochester 1979. Essays on rational expectations and flexible exchange rates

WALTER SCHUBERT, Ph.D. Rutgers 1978. The effects of foreign direct investment on the level of human capital employed in the Southeast.

WILLIAM R SCHULTZ, Ph.D Iowa State 1979 The impact of short-term capital flows and balance-of-payments fluctuations on monetary policy An empirical study of Japan.

BRUCE M. SEIFFERT, Ph D Michigan 1978. The individual's demand for assets denominated in various currencies.

JOSÉ M. SOLÉ-MARIÑO, Ph D Northwestern 1978. Balance of payments, output, and prices The experience of Spain, 1960–75.

WOONKI SUNG, Ph D Michigan 1979 The Euro-dollar market: A study of its growth

AMOR TAHARI, Ph.D Pittsburgh 1978 Money, government budget constraint, and balance of payments in a growing economy.

STEPHEN S. THURMAN, Ph.D. George Washington 1979 International price linkages: An analysis from the standpoint of a large econometric model of the *U S* economy.

ABDUL M. TURAY, Ph.D Oklahoma 1978. The EEC and the *LDC*'s: Trade association agreements with special emphasis on the Lomé convention.

EVAN S. URDANG, Ph.D. Pennsylvania 1979 An international flow-of-funds model with an endogenous exchange rate.

JAN VANOUS, Ph.D. Yale 1979. An econometric model of world trade of member countries of the Council for Mutual Economic Assistance.

BRENT D. WILSON, D.B.A. Harvard 1979. The disinvestment of foreign subsidiaries by *U.S.* multinational companies.

THOMAS S. WURSTER, Ph.D. Yale 1978. The firm in the international environment.

CHARLES A. WYPLOSZ, Ph.D. Harvard 1978. Three theories of the trade balance: A case study of France.

YUNGHO YOO, Ph.D. Wisconsin (Madison) 1978. The Heckscher-Ohlin theorem and *U.S.* trade in manufacturing commodities.

SOOGIL YOUNG, Ph.D. Johns Hopkins 1979. Inflation, output, and the balance of payments under fixed exchange rates with long-term automatic adjustment: A reconciliation of the Keynesian and the monetarist analyses.

SALEHA ZHOWANDAI, Ph.D. Hawaii 1978. An economic analysis and measurement of Afghanistan protection structure.

RICHARD A. ZUBER, Ph.D. Kentucky 1978 A model for constructing efficient currency cocktails.

Business Administration; including Business Finance and Investment, Insurance, Marketing, and Accounting

MANOJ K. AGARWAL, Ph.D. State University of New York (Buffalo) 1978. Prediction of consumer choice and demand forecasting: An evaluation of alternative approaches.

ROLF W. BANZ, Ph.D. Chicago 1978. Limited diversification and market equilibrium: An empirical analysis

CARL BATLIN, Ph.D Columbia 1978 Asset leasing under conditions of costly bankruptcy

ARTHUR D. BEARD, Ph.D. Chicago 1978. The quantification of visual esthetic judgment using verbal and nonverbal scales

CONSTANT E. BECKERS, Ph D. California (Berkeley) 1979. Empirical studies of alternative option pricing models.

MICHEL J. BERGIER, Ph.D. State University of New York (Buffalo) 1978 An examination of consumer's leisure time choice behavior: A multidimensional scaling approach to the study of spectator sports

THOMAS V. BINGHAM, Ph D. Cincinnati 1978 Economic models: The large multiproduct firm and its organizational structure.

JAMES H. BITEMAN, D.B.A. Harvard 1979 Turnaround management: An exploratory study of rapid, total organization change

ESER BORAK, Ph.D. Columbia 1978 A model of dyadic interaction between salesman and prospect in the life insurance industry in Turkey.

IVAN BRICK, Ph.D. Columbia 1979 The debt maturity structure decision.

PETER BROWNELL, Ph.D. California (Berkeley) 1979. Participation in budgeting, locus of control, and organizational effectiveness.

JAMES A. BUSZUWSKI, Ph.D. Boston College 1979. An applied micro-economic comparison of three theories of investment behavior.

EDWARD G. CALE, JR., D.B.A. Harvard 1979. Implementing standard computer-based application programs in multiple site organizations.

DIPANKAR CHAKRAVARTI, Ph.D. Carnegie-Mellon 1979. A cognitive approach to the development and implementation of marketing decision models.

JAMES G. CLAWSON, D.B.A. Harvard 1979. Superior subordinate relationships in managerial development.

JEFFORD R. CORNWALL, Ph.D. Michigan 1979. Leadership, personality, and organizational performance.

MARCEL L. CORSTJENS, Ph.D. California (Berkeley) 1978. Lancor: An idiosyncratic deterministic continuous choice model.

LAWRENCE A. CROSBY, Ph.D. Michigan 1979. Cognitive effects of consumer information and education.

HARRY DAVIS, Ph.D. Columbia 1978. The growth margin ratio of fifo and lifo firms: Implications for operationalizing the Trueblood criterion and insight into price adjustment lags.

PRADUDDHA DE, Ph.D. Carnegie-Mellon 1979. Essays on database and scheduling systems.

DAVID F DE ROSA, Ph.D. Chicago 1978. Rates of return on common stocks and inflation: Theories and tests.

GIDEON FALK, Ph D Chicago 1978. An examination of some normative effects of unanimity and majority rules on the quality of solutions in problem solving groups with unequal power.

KAREN N. GAERTNER, Ph.D Chicago 1978. Organizational careers in public school administration.

MORTON GALPER, D B A Harvard 1979. Improving communications spending decisions for industrial products

CARMELO GIACCOTTO, Ph D Kentucky 1978. The random walk hypothesis of stock market prices reconsidered

VIJAYARAGHAVAN GOVINDARAJAN, D.B.A. Harvard 1978 Behavioral effects of audits in the health care services A study in the field using exploratory data analysis and classical statistical techniques.

LEI L GREMILLION III, D.B A Harvard 1979. Managing the organizational learning process in implementing computer-based systems

ROBERT GRUBER, Ph.D. Columbia 1978. The impact of response set on large-scale survey research involving activities, interests, and opinions (AIO) variables.

LOUIS M. GUTENTAG, D.B A. Harvard 1979. Acceptance of computer-based medical consultation systems: A study of physician attitudes

KATHRYN R. HARRIGAN, D.B A Harvard 1979. Strategies for declining businesses.

RANDALL B. HAYES, Ph D. Michigan 1978. The information content of replacement cost data.

ROBERT L. HEINKEL, Ph.D. California (Berkeley) 1978. Essays on financial markets with asymmetric information

PHILIP E. HENDRIX, Ph.D. Michigan 1978. The allocation of time by consumers.

EDWARD E. HOSKINS, Ph.D. Mississippi 1976. A reha-

bilative effect upon debtors of chapter XIII of the U.S. Bankruptcy Act.

TAIWO D. IDEMUDIA, Ph.D. State University of New York (Buffalo) 1978. An inquiry into the performance of Nigerian indigenous entrepreneurs.

FRANKLIN J. INGRAM, Ph.D. South Carolina 1979. A multivariate test for discrimination in the residential mortgage lending patterns of financial institutions

HAJI Y. IZAN, Ph.D. Chicago 1978. An empirical analysis of the economic effects of mandatory audit regulation.

V. JAIKUMAR, Ph.D. Columbia 1979. An application of mathematical control theory to the accounting model for management control.

KEM R. KEMATHE, Ph.D. Pittsburgh 1978. Selected characteristics of black businesses in the United States, 1969–72, with two case studies of black-owned businesses in Pittsburgh as special reference

RAFAEL LAZIMY, D B A. Harvard 1978 Interactive preference assessment and constrained optimization with multiple conflicting objectives

FRANK S. LEONARD, D.B.A. Harvard 1978. Managerial assumptions: The underlying structure for managerial behavior in the U.S. Forest Service.

EDWARD LEWIS, Ph.D. State University of New York (Buffalo) 1978. An algorithm for budget allocation, recharge, pricing, and adaptive control for computer resources.

JOSHUA LIEBERMANN, Ph D. Chicago 1978 A theory of price deals in supermarkets

DAVID T. LIVINGSTON, Ph D South Carolina 1978. The relationship between employee ownership and financial characteristics of the firm An empirical investigation

NARESH K MALHOTRA, Ph D State University of New York (Buffalo) 1979. Consumer information seeking and information processing. An alternative theoretical framework and some empirical investigations.

CARL W. H MARGRABE, Ph.D Chicago 1978. The value of farm real estate in the United States, 1912–75.

STEVEN E. MARKHAM, Ph.D State University of New York (Buffalo) 1979. Leadership and motivation: An empirical examination of exchange and its outcomes.

PETER F MATHIAS, D.B.A. Harvard 1979 The role of the logistics system in strategic change· The case of the multinational corporation

KENNETH A. MERCHANT, Ph.D. California (Berkeley) 1978. Departmental budgeting An empirical examination of a contingency model.

JOSEPH M. MESSINA, Ph.D. California (Berkeley) 1979. An empirical investigation of a multiperiod portfolio model of the term structure of interest rates.

WILLIAM C. NEMITZ, Ph.D. Boston College 1979 Manufacturing investment: A three-component expectational model.

STEPHEN H. PENMAN, Ph.D. Chicago 1978. Corporate forecast disclosure, substitute information, and the market for information.

JANIS R. REEDER, Ph.D. South Carolina 1978. The child care allowance issues in horizontal equity.

KAMI S. P. RWEGASIRA, Ph.D. Columbia 1979. The application of a credit strength index to improve financial analysis in Tanzanian commercial bank lending.

DAN SAREL, D.B.A. Harvard 1978 An investigation of the effect of risky attributes on consumer evaluation strategies.

LESLIE L. SPERO, D.B A. Harvard 1979. The extent and causes of voluntary disclosure of financial information in three European capital markets: An exploratory study.

MARILYN L. TAYLOR, D.B.A. Harvard 1979 The role of the staff specialist in the policy implementation process.

WILLIAM M. TAYLOR, Ph.D. Chicago 1979. Stamps as an investment An analysis of a price series with quality variation.

HASSAN TEHRANIAN, Ph.D. Alabama 1979. Empirical studies in portfolio performance using stochastic dominance

ROLAND VAN DIERDONCK, D B A Harvard 1979. The materials management process· A conceptual framework

CHARLES M. WADE, Ph.D Mississippi 1977. A capital equipment budgeting system applicable to selected mid-South not-for-profit community hospitals

HAROLDENE F. WUNDER, Ph D South Carolina 1978 Capital formation and the investment credit: An empirical study

DOUGLAS K. ZIMMERMAN, Ph D State University of New York (Buffalo) 1979. Nurse-doctorates Patterns of role orientation in a professional elite.

THOMAS S ZORN, Ph D California (Los Angeles) 1978 Information differences, the stock market, and managerial incentives

Industrial Organization and Public Policy; including Economics of Technological Change, and Industry Studies

MICHAEL ADAMS, Ph.D. California (Santa Barbara) 1977. An interregional price and allocation analysis of the *U.S.* softwood plywood industry.

PETER ADELSHEIM, Ph D. American 1979 Structural characteristics of the firm and the market: The impact on power and performance

SHIRLEY C. ANDERSON, Ph D. California (Los Angeles) 1979. The economic effects on urban bus transit of public ownership and urban mass transportation administration capital grants.

LEWIS A. ASSALEY, Ph.D. Cincinnati 1979. Police productivity: A value and cost analysis.

JAY M ATKINSON, Ph.D. Virginia Polytechnic Institute 1978. A study of regulatory goals and controls: Firm size in the savings and loan industry.

DIANA BARRETT, D.B.A. Harvard 1979. Multihospital systems: The process of development.

BARRY D. BAYSINGER, Ph.D Virginia Polytechnic Institute 1978. A theory of the efficiency of jurisdictional choice: The case of corporate federalism.

JAMES P. BELL, Ph D. California (Berkeley) 1978. Class price discrimination under regulation. Electric utilities.

Y. BELLICHA, Ph.D. McGill 1979 Estimating returns to scale in selected manufacturing industries in Canada.

MICHAEL S BERMISH, Ph.D. City (New York) 1979. A short-run forecasting model of the market for carpets.

LESLIE L. BODEN, Ph.D. Massachusetts Institute of Technology 1977. Underground coal mining accidents and government enforcement of safety regulations

GERALD R. BODISCH, Ph.D Virginia 1979. Market power and employment fluctuation.

PERRY L. BROWN, Ph.D Minnesota 1978 Is collusion observationally identifiable?

THOMAS H. BRUGGINK, Ph.D Illinois 1979 Efficient allocation of municipal water. Comparison of regulated private enterprise with public enterprise

JAMES E. BURI, Ph.D. Notre Dame 1979 The role of the regulations of the Securities and Exchange Commission as a means to strengthen federal policy regarding conglomerate mergers.

ROBERT D CAIRNS, Ph D. Massachusetts Institute of Technology 1978. Rents in the Canadian nickel industry.

BALAJI S. CHAKRAVARTHY, D.B A. Harvard 1978. Adapting to changes in the coal industry: A managerial perspective

DON CLARK, Ph D Michigan State 1978. The structure of transportation costs in international trade

LINDA COHEN, Ph D. California Institute of Technology 1979. Essays on the economics of licensing nuclear power plants.

JOHN D. CULBERTSON, Ph D. Wisconsin (Madison) 1979. Firm invention and market power in the food industries.

JOSEPH R. D'CRUZ, D B A Harvard 1979 Quasi integration in raw material markets. The overseas procurement of coking coal by the Japanese steel industry

ALLAN S. DETSKY, Ph D. Massachusetts Institute of Technology 1978 An economic analysis of policy issues in the supply of medical services

VAUGHN DICKSON, Ph.D. Western Ontario 1977. Scale efficiency in Canadian manufacturing.

FORREST P. DIXON, Ph.D. Illinois 1979 Implications of joint venture contracting on the structure of the Pacific Northwest beef industry.

JERRY DUVALL, Ph.D. American 1979. The induced innovation hypothesis and biased technological change in the domestic telecommunications industry: A micro-econometric analysis.

DAVID M. EISENSTADT, Ph.D. Illinois 1979. The role of Blue Shield in the health service sector: A theoretical and empirical analysis.

AHMAD FARUQUI, Ph.D. California (Davis) 1979. The dynamics of energy substitution in California manufacturing, 1958–76.

WENDY GARLING, Ph.D. American 1979. The steam car market in the United States: A quantitative analysis of short-run supply and demand.

AMIHAI GLAZER, Ph.D. Yale 1978. The economics of repairs.

ANDREW A. GOETT, Ph.D. California (Davis) 1978. An application of quantal models to household energy decisions.

EARL L GRINOLS, Ph.D. Massachusetts Institute of Technology 1977. Imperfect markets and government behavior

DANIEL M HAMBLIN, Ph.D. State University of New York (Buffalo) 1979. Automobile demand models for evaluating impacts of fuel efficiency energy policies.

JAMES T HANNON, Ph.D. Illinois 1978 The impact of competition on cost allocation in a multimarket monopoly· Telecommunications.

HENRY HANSMANN, Ph.D Yale 1978. Three essays on the role of nonprofit enterprise.

MARC W HEROLD, Ph D. California (Berkeley) 1979. The determinants of American investment in Latin American manufacturing.

JOHN C. HILKE, Ph D Cornell 1978. Determinants of entry. Cost and risk barriers to entry.

RAY C. HUTTSELL, JR., Ph.D. Nebraska (Lincoln) 1979. The impact of entry regulation by the Nebraska Public Service Commission upon the motor carrier industry in Nebraska.

RICHARD P. HYDELL, Ph.D. Massachusetts Institute of Technology 1978. A study of technological diffusion: The replacement of steam by diesel locomotives in the United States

WILLIAM F IRISH, Ph.D North Carolina State 1978. An econometric analysis of seasonal elasticities, effects of alternative price specifications, and long-run elasticities for forecasting commercial KWH sales.

ALI D. JOHANY, Ph.D. California (Santa Barbara) 1978. OPEC is not a cartel: A property rights explanation of the rise in oil prices.

NEIL JOHNSON, Ph.D. Johns Hopkins 1979. An evaluation of the paradox of short-run increasing returns to labor with empirical analysis for Canadian manufacturing industries.

STEPHEN A. JOHNSTON, Ph.D. North Carolina State 1979. Construction work in progress in the rate base vs. allowance for funds used during construction: An investigation of the effects on capital intensity in electric utilities.

RUSSELL O. JONES, Ph.D. California (Santa Barbara) 1977. Vertical integration, cartel coordination, and the petroleum industry.

SURAJ KANHOUWA, Ph.D. American 1979. The feasibility of financing *U.S.* energy development during 1976–85.

KAMBIZ H. KIANI, Ph.D. State University of New York (Binghamton) 1979. An empirical study of optimal plant flexibility.

WILLIAM W. KIFFMEYER, Ph.D. Wisconsin (Milwaukee) 1978. The determination of prices in *U.S.* manufacturing industries.

GEORGE J. KOWALSKI, Ph.D. Queen's (Kingston) 1978. Pricing, entry, and performance in the nickel industry.

ANNETTE LAMOND, Ph.D. Yale 1979. Competition in the general freight motor carrier industry.

JANET LANDA, Ph.D. Virginia Polytechnic Institute 1978. The economics of the ethnically homogeneous Chinese middleman group: A property rights-public choice approach.

ROBIN C. LANDIS, Ph.D. Massachusetts Institute of Technology 1979 Sulfur emissions policies, oil prices, and the Appalachian coal industry.

WALTER LANE, Ph.D. California (San Diego) 1979. Product differentiation in a market with endogenous sequential entry.

TOM KWAN-YAU LEE, Ph.D. California Institute of Technology 1978. Micro-economic foundations of research and development.

KENNETH LEESON, Ph.D. New York 1979. A cohort analysis of costs, prices, sales, and profits in book publishing.

ANDRÉ LEMELIN, Ph.D. Harvard 1979 Patterns of inter-industry diversification by Canadian manufacturing industries.

SAUL LEVMORE, Ph.D. Yale 1978. Small firms' disadvantages in television advertising.

STANLEY J. LIEBOWITZ, Ph.D. California (Los Angeles) 1978. Concentration and profitability: A look at some alternative explanations.

HENRY BOON FUI LIM, Ph.D. Chicago 1979. Cost-benefit analysis for peak load pricing and load management.

LOUIS LOPILATO, Ph.D. Fordham 1979 An analysis of expansion in the American steel industry, 1978–83.

MICHAEL M. LUCKEY, Ph.D. Michigan 1978. An econometric model of the demand for automobiles in the United States.

DAVID W. MCLAMB, Ph.D. North Carolina State 1978. Industrial demand for city-supplied water and waste treatment.

ANDREW J. MCLAUGHLIN, Ph D. California (Los Angeles) 1979. An economic analysis of resale price maintenance.

GARY P. MAHAN, Ph.D. North Carolina State 1978. The demand for residential local and long-distance telephone service: An analysis of household data.

ROBERT W. MANDELL, Ph.D. Fordham 1979. The financing and capital needs of the domestic airline industry in the 1980's.

PAUL MARCHETTI, Ph.D. Maryland 1979. Consumer preferences and industry competition: A simulation study.

MARK MEINERS, Ph.D. Georgetown 1978. Nursing home costs: A statistical cost function analysis using national survey data.

ILONA MELSTRADS, Ph.D George Washington 1979. Patentability of living organisms.

ROBERT O. MENDELSOHN, Ph.D. Yale 1978. Towards efficient regulation of air pollution from coal-fired power plants.

GALE D. MERSETH, D.B.A. Harvard 1979. Regulation and income redistribution· The Massachusetts experience.

MICHAEL R. METZGER, Ph.D. Stanford 1978. A theory of monopolistic competition with imperfect consumer information: A treatment of inflation.

LARRY J. MOFFITT, Ph.D. California (Berkeley) 1979. A disequilibrium analysis of agricultural, commercial, and consumer gasoline demand.

DEAN MOUNTAIN, Ph.D. Western Ontario 1979. Inter-fuel substitution in the electricity industry.

MUSTAFA K. MUJERI, Ph.D. McMaster 1978. The world market for jute: An econometric analysis.

ARUN K. MUKHOPADHYAY, Ph.D Brown 1979 Diversification, research productivity, research intensity, and market shares.

TOSHIAKI NAKAZAWA, Ph.D Wisconsin (Madison) 1979. Competition among firms as revealed in market share fluctuations

RANDY A. NELSON, Ph.D. Illinois 1978. Regulation and technical change in the electric utility industry.

JOSEP OLIU, Ph.D. Minnesota 1978. Peak load with "too-high" metering costs· Optimal pricing under myopia.

CLIFTON V. OLSON, Ph D Tulane 1979. Policy objectives in airline regulation: An analysis of the performance of the CAB.

ARIEL S. PAKES, Ph.D. Harvard 1978. An empirical analysis of the role of economic incentives in the production and transmission of economically useful knowledge.

PAUL S. PEYSER, Ph.D. Wisconsin (Madison) 1979 Merger and capital structure adjusted changes in shareholder welfare.

MAHESH K. PODAR, Ph D. Pennsylvania State 1978. A benefit-cost analysis of fuel economy standards for passenger cars, model years 1981–84.

NAGABHUSHA PULIKONDA, Ph D. Illinois 1977. The impact of vertical integration on regulatory constraints: Theory and testing.

JAMES RADAR, Ph.D. State University of New York (Buffalo) 1979. An analysis of the strategies employed by German and Japanese auto manufacturers in penetrating the *U.S.* automobile market during the years 1965–76.

FARSHAD RAFII, D.B.A. Harvard 1978. Joint ventures and transfer of technology to Iran: The impact of foreign control.

SITA RAM MOHAN, Ph.D. Purdue 1978. Public utility regulation in a changing environment

EDWARD M. RICE, Ph.D. California (Los Angeles) 1978 An examination of interindustry price flexibility differences.

EDWARD J. RIFKIN, Ph.D. California (Santa Barbara) 1978. Dynamic limit price theory and market growth.

TANYA ROBERTS, Ph.D. Washington 1979. Federal price regulation of close substitutes for fresh drinking milk: History, economic analysis, and welfare implications.

CLIFFORD B. ROCHLIN, Ph.D. California (Santa Barbara) 1979. Essays in advertising.

VALERIE SARRIS, Ph.D. Yale 1977. The efficiency of private antitrust enforcement: The Illinois brick decision and its implications.

EDWARD L. SATTLER, Ph D. Illinois 1978. Conglomerate mergers and the railroads.

RONALD S. SAUNDERS, Ph.D Harvard 1978 The determinants of the productivity of Canadian manufacturing industries relative to that of counterpart industries in the United States

ROY T. SAVOIAN, Ph.D California (Santa Barbara) 1979 Horizontal merger and the efficacy of divestiture: An empirical approach

MARK SCHAEFER, Ph.D. Massachusetts Institute of Technology 1977 Three essays on the economics of congested service facilities.

MARK A. SCHANKERMAN, Ph D Harvard 1979. Essays on the economics of technical change: The determinants, rate of return, and productivity impact of research and development.

FREDERICK J. SCHROEDER, Ph D. Wisconsin (Madison) 1978. Joint production of externalities An empirical analysis of cost and production structure and waste treatment demand in the *U.S.* brewing industry

MARK A. SCHWARTZ, Ph.D Pennsylvania 1978 The imitation and diffusion of industrial innovations.

FRANK A. SCOTT, Ph.D. Virginia 1979. An economic analysis of fuel adjustment clauses

JOSEPH SHAANAN, Ph.D Cornell 1979. A test of limit pricing: An empirical study of entrant and existing firm behavior.

MANESH L SHRIKANT, D.B.A. Harvard 1979 Managing the external competitive context: A management function in the regulated environment of India.

ALBERT K. SMILEY, Ph.D. Princeton 1978. Competitive bidding under uncertainty The case of offshore oil.

JOHN J. SNYDER, Ph.D. Colorado 1979. Residential electricity demand in Colorado municipalities: A time-series cross-section study.

PAUL E. SOMMERS, Ph.D. Yale 1978. The diffusion of nuclear power generation in the United States.

RICHARD H. SPADY, Ph.D. Massachusetts Institute of Technology 1978. Econometric estimation of cost functions for the regulated transportation industries.

MATTHEW L. SPITZER, Ph.D. California Institute of Technology 1979. Axiomatic analysis of legal/institutional issues.

DAVID C. STAPLETON, Ph.D. Wisconsin (Madison) 1978. Nonprice rationing in the demand for medical education.

MEIR STATMAN, Ph.D. Columbia 1979. The measurement of monopoly power: A method and application to the U.S. ethical pharmaceutical industry.

PANKAJ TANDON, Ph.D. Harvard 1979. Aspects of optimal *R&D* policy.

MICHAEL TANSEY, Ph.D. Wisconsin (Madison) 1978. Output effects of price controls: The economic stabilization program, 1971–74.

LACY G THOMAS III, Ph.D. Duke 1979. Essays on the regulation of consumer product safety.

ALEXANDER M. THOMPSON III, Ph D. Stanford 1979. Capital accumulation in *U.S.* coal: A case study of labor, technology, and competition.

SOEHFROE TJOKROPRAJITNO, Ph.D. Washington State 1979. The economic impacts of subsidy programs on the level of *US* participation in cargo liner trade.

JAMES C VERTREES, Ph D. California (Berkeley) 1979. The cost of medical residency program in an outpatient setting.

MARK A VONDEREMBSE, Ph.D. Michigan 1979. An investigation of design and operating decisions for continuous casting in the steel industry.

BARBARA A. WALDMAN, Ph.D New Mexico 1978. A cost-benefit analysis of accuracy in peak load forecasting.

DAVID H. WATERMAN, JR., Ph.D Stanford 1979. Economic essays on the theatrical motion picture industry.

JAMES W H. WATSON, Ph.D. California (Los Angeles) 1978. The effects of Federal Energy Administration regulations and dealer-day-in-court laws on service at gasoline stations.

STEVEN N WIGGINS, Ph.D. Massachusetts Institute of Technology 1979. Product quality regulation and innovation in the pharmaceutical industry.

MARK WILLIS, Ph.D. Yale 1979. The effects of cyclical demand on industry structure and on the rate of technological change: An international comparison of the housebuilding sectors in the United States, Great Britain, and France.

RONALD G. WIRICK, Ph.D. Western Ontario 1978. Price supports vs. direct subsidies: A simulation test of the federal dairy support program, 1969–72.

KENNETH T. WISE, Ph.D. Massachusetts Institute of Technology 1979. The effects of OSHA regulations on the *U.S.* lead industry: An economic impact and economic modeling analysis.

BENJAMIN T. YU, Ph.D. Washington 1978. The economics of patent pools: The capturing of the return to basic research.

Agriculture and Natural Resources

MOHAMMED H. ABDALLAH, Ph.D. Ohio State 1978. Economics of corn stover as a coal supplement in steam-electric power plants in the North Central United States.

MOHAMMED I. AHMAD, Ph.D Washington State 1979. An economic analysis of rate sensitivity to loan structure component variations within the U.S banks for cooperatives using a multiperiod linear programming approach.

ORLEY M. AMOS, JR, Ph.D. Iowa State 1979 Supply of potential cropland in Iowa.

GEIR B. ASHEIM, Ph.D. California (Santa Barbara) 1979. Renewable resources and paradoxical consumption behavior.

MARCELINO AVILA, Ph.D Missouri (Columbia) 1979. An economic evaluation of alternative annual cropping systems in two regions of Costa Rica.

WILLIAM C. BAILEY, Ph.D. Missouri (Columbia) 1978 Optimum export facility location for least cost air transportation of livestock exported from the United States.

HARRY S. BAUMES, JR, Ph.D Purdue 1978 A partial equilibrium sector model of *U.S* agriculture open to trade. A domestic agricultural and agricultural trade policy analysis.

JOSEPH C. CAMPBELL, Ph.D. Iowa State 1978 Agricultural exports and sediment effluent: Their relationship, and the economic and environmental impacts of national sediment control policies.

CONSTANCIO C CANETE, Ph.D Purdue 1978 The income level and income distribution impacts of the Masagana 99 Program in Central Luzon, Philippines.

MAN-WAH LUKE CHAN, Ph D. McMaster 1978 A model of the Canadian beef and dairy cattle industry based on Markov chain techniques.

JEAN-PAUL CHAVAS, Ph.D. Missouri (Columbia) 1978 A quarterly econometric model of the *U.S* poultry and egg industry.

RICHARD A. CLARK, Ph.D. Cornell 1979. The optimal allocation of federal coal land among competing uses.

RODNEY L. CLOUSER, Ph.D Purdue 1979. Incorporation of water-intensive activities into estimates for rural household water demand.

ROBERT CONRAD, Ph.D. Wisconsin (Madison) 1978. *Taxation and the theory of the mine.*

GREGORY S. CRESPI, Ph.D. Iowa 1978. The economics of artificial upwelling mariculture: A preliminary assessment.

ABDULLA DANOK, Ph.D. Purdue 1979. Effect of weather variability on machinery selection.

JOHN T. DAUBERT, Ph.D Colorado State 1979. The economics of low-flow equations.

SUPOTE DECHATES, Ph.D. Washington State 1978 On-farm impacts of economic policy choices in Thailand.

YAHAYA DOKA, Ph.D. Washington State 1979. Policy objectives, land tenure, and settlement performance *Implications for equity and economic efficiency* in the Columbia Basin irrigation project.

ANDREW K. DRAGUN, Ph.D. Wisconsin (Madison) 1978. Economic perspectives of environmental quality: A comparative analysis of three divergent views.

RICHARD L. DROBNICK, Ph.D. Southern California 1979 Food policy choices in an uncertain world: Policy analysis with a probabilistic world food model.

DANIEL J. DUNN, Ph.D Illinois 1978 Lending risks of the federal land banks An application of the loan performance simulator.

WILLIAM M. EDWARDS, Ph D Iowa State 1979 Farm machinery selection under conditions of uncertainty.

ALI ERYILMAZ, Ph D California (Davis) 1978 An econometric analysis of the California apricot industry

EDMUND A ESTES, Ph D. Washington State 1979 Supply response and simulation of supply and demand for the *U.S* potato industry

HAL EVERETT II, Ph.D. Purdue 1979 Changing energy supply and demand in the *U.S* economy· A demonstration model

CLAUDE F. FALCON, Ph D. Michigan State 1978. Preretirement exit of farm operators. Conceptual framework and empirical study

CHRISTOPHER F FEISE, Ph D. Washington State 1978. A radical analysis of the transformation of American agriculture and its impact on labor, 1910–75.

ARAY M FELDENS, Ph.D Ohio State 1978 A transportation-storage network analysis of wheat and soybeans in Rio Grande Do Sul, Brazil

JOSE LUIS FERNANDEZ-CAVADA, Ph D California (Davis) 1979 International trade in fresh oranges and tangerines. Analysis of potential structural changes including EC expansion

KENNETH FLEMING, Ph.D. California (San Diego) 1979 An econometric analysis of selected Air Force quality standards

JOHN M. FOWLER, Ph.D. Iowa State 1979 The interface of forestry and agriculture as nonpoint sources of suspended sediment: A national modeling approach

MARIA C. GALLASTEGUI, Ph.D. Brown 1979. The exploitation of a renewable and common property resource· The case of sardine fishing on the Mediterranean coast of Spain.

STEVEN G. GOLD, Ph.D State University of New York (Binghamton) 1979. The economic impact of instituting effluent charges for the management of air quality in the *U.S.* electric power industry

DOUGLAS A. GREENLEY, Ph.D. Colorado State 1979. Recreation and preservation benefits from water quality improvement.

BRUCE F. HALL, Ph.D California (Berkeley) 1979 An economic analysis of multiple job holding by farm operators.

MICHAEL C. HALSE, D.B A Harvard 1979. The concept of conversion efficiency as applied to the Indian milk and food economy.

MICHAEL D. HAMMIG, Ph D. Washington State 1978 Supply response and simulation of supply and demand for the *U.S* fresh vegetable industry

W MICHAEL HANEMANN, Ph.D Harvard 1978. A methodological and empirical study of the recreation benefits from water quality improvement

LARS P HANSEN, Ph.D. Minnesota 1978 Econometric modeling strategies for exhaustible resource markets with applications to nonferrous metals

AHMED HASHEMI, Ph D City (New York) 1978 The international crude oil prices, 1957-76. An exhaustible resource and an imperfect competition

DONALD J HERTZMARK, Ph D Wisconsin (Madison) 1978. Economics, thermodynamics, and the allocation of energy.

ISMAIL A HUMMADI, Ph D Colorado 1978. Economic growth and structural changes in the Iraqi economy with emphasis on agriculture, 1953 75

CATHY L JABARA, Ph D Purdue 1979 Agricultural comparative advantage under uncertainty The case of Senegal

LAURENCE R JACOBSEN, Ph D Stanford 1979 Optimal investment and extraction rates for a depletable resource

SURAPAN JANJAROEN, Ph D Purdue 1979 An economic analysis of recent developments in Pacific Basin corn trade

WATTANA S JANJAROEN, Ph D Purdue 1979 Cost sharing of water project, game theory approach A case study of the Chao Phya Irrigation Improvement Project, Stage II, Thailand

HEZEKIAH S JONES, Ph D Illinois 1979 Solar energy as an alternative to fossil fuel for drying corn on farms in east central Illinois

ROBERT J JONES, Ph D Columbia 1978 A capital accumulation analysis of intercountry differences in education and economic growth

JOHNNY W JORDAN, Ph D Clemson 1978 The effect of changing tobacco harvest machinery technology on resource allocation and product adjustment in the upper coastal plains of South Carolina

RYOHEI KADA, Ph.D. Wisconsin (Madison) 1978 Off-farm employment and farm adjustments. Microeconomic study of the part-time farm family in the United States and Japan

JAMES N. KIMBALL, Ph.D. Virginia Polytechnic Institute 1979. The economic effect of water effluent regulation and enforcement upon firm behavior.

HENDRIK C. KNIPSCHEER, Ph.D Illinois 1979. An econometric model of the demand for soybeans and soybean meal in the countries of the European Economic Community.

DAVID M. KOHL, Ph.D. Cornell 1978. Merchant and dealer credit in New York State.

YODYING KONGTONG, Ph.D. Iowa State 1978. Use of subjective data in estimating farm supply response.

EUGENE A. KROCH, Ph.D. Harvard 1978. Dynamics of energy resource allocation in theory and in practice: Modelling nuclear power alternatives.

N JOHN KURIAN, Ph.D. Western Ontario 1977. Behavior under uncertainty and public policies in peasant agriculture: Applications of economic theory of uncertainty.

EDGAR A LANZER, Ph D California (Davis) 1978. Fertilizer recommendations from the dynamic Leibig-Mitscherlich model: The case of wheat-soybeans in southern Brazil

ALFRED L LEVINSON, Ph D. California (Berkeley) 1979. Energy and materials in three sectors of the economy: A dynamic model with technological change as an endogenous variable

ANDRES M LIEBENTHAL, Ph.D West Virginia 1978 Interstate supply competition in Appalachian steam-coal markets

FU-SHAN LIU, Ph D Illinois 1979. Interregional flows of slaughter hogs and pork carcasses in Taiwan.

VERNE W. LOOSE, Ph.D British Columbia 1978. Optimal exploitation of a salmon fishery: A simulation approach

DAVID A MCKINNA, Ph D. Cornell 1978. Agricultural export market development procedures and practices with special emphasis on *U S* government-sponsored programs

RODOLFO d M MATIENZO, Ph D Ohio State 1978. Repayment and group lending in the province of Camarines Sur, Philippines, 1976-77.

ABDUL J MIAN, Ph D. Ohio State 1978 A least-cost nutritional diet for Bangladesh: An alternative approach to food supply.

JOSEPH W MILON, Ph D Florida State 1978. Evaluating economic and thermodynamic policy models: A study of the effect of residential solar water heating on electric utilities in Florida.

DONALD M MOLIVER, Ph.D. Virginia Polytechnic Institute 1978. Oil and money in Saudi Arabia.

ALFONSO T. MONARDES, Ph.D Cornell 1978. An economic analysis of employment in small farm agriculture: The central valley of Chile.

MICHAEL J MUELLER, Ph.D. Michigan 1978. Oil shale, economic rent, and public policy.

ROBERT D. NIEHAUS, Ph.D. Maryland 1979. The social cost of stratospheric modification from agriculture nitrogen use.

JAY E. NOEL, Ph.D. California (Davis) 1979. Dynamic water management. An optimal control approach.

GEORGE W. NORTON, Ph.D. Minnesota 1979 A model

for Indian Reservation agricultural development: The case of the Sisseton-Wahpeton Sioux.

ANDREW M. NOVAKOVIC, Ph.D. Purdue 1979. An economic analysis of the *U.S.* dairy price support program and alternate policies.

ETHELBERT C. NWAGBO, Ph.D. Illinois 1979. Optimal responses to market risk by farmer cattle feeders.

KEIJIRO OTSUKA, Ph.D. Chicago 1979. Allocation of public to agricultural research: Japan 1950–75 with special reference to the pricing of rice.

PAUL T. PERRAULT, Ph.D. Stanford 1979. Banana-manioc farming systems of the tropical forest: A case study of Zaire.

DENNIS W. PERVIS, Ph.D. Michigan State 1979. Simulation of a parastatal fertilizer marketing with reference to Columbia: An application of the general system simulation approach.

ABDUHL RADHI, Ph.D. Purdue 1979. An economic model of the world nitrogen fertilizer market

DARYLL D. RAITT, Ph.D. Missouri (Columbia) 1978. An economic analysis of potential water resource development for agriculture in the Blackwater-Lamine River Basin, Missouri.

NIMAL F. C. RANAWEERA, Ph.D. California (Davis) 1978. Land settlement in Sri Lanka: The Mahaweli Ganga Project.

MICHAEL R. REED, Ph.D. Iowa State 1979. Feed grain imports and their effect on feed grain prices in the importing country.

JOSE L. RIBEIRO, Ph.D. Minnesota 1979. Rates of return to agricultural investment in the Cerrados area in Brazil.

CHRISTOPHE J. RIBOUD, Ph.D. Massachusetts Institute of Technology 1978. Welfare effects of farm commodity programs in the United States: 1948–73.

CHARLES H. RIEMENSCHNEIDER, Ph D. Michigan State 1978. An information systems analysis of USDA farm income data.

CLARENCE D. ROGERS, Ph.D. Clemson 1978. An economic evaluation of cotton consumption related to quality characteristics.

FRANK S. ROSE, Ph.D Michigan State 1979. Economic analysis of alternative dairy import policy for the United States.

RONALD F. ROST, Ph.D. Northwestern 1978. Pollution controls and labor productivity growth.

TIMOTHY J. RYAN, Ph.D. Minnesota 1978. Commodity price determination and transmission: An analysis of the farm-retail pricing of *U.S.* beef.

LARRY D. SCHELHORSE, Ph.D. California (Santa Barbara) 1979. Commercial demand for electricity: Evidence for a single utility.

JOHAN SCHOLVINCK, Ph.D. Cornell 1979. Migration and population growth in rural areas of New York State.

HAIM SHALIT, Ph.D. California (Berkeley) 1979. Analysis of increasing farmland values.

RICHARD C. SHANE, Ph.D. Washington State 1978. An interregional analysis of the effects of increased potato production, processing and transportation factor-input costs on the *U.S.* potato industry.

HOSSEIN SHAPOURI, Ph.D. Washington State 1979. The supply and demand for the *U.S.* livestock and broiler sectors.

SHAHLA R. SHAPOURI, Ph.D. Washington State 1979. The demand for *U.S.* wine.

MICHAEL F. SHEEHAN, Ph.D. California (Riverside) 1979. The theory of the limiting utility applied to the development of natural resources in arid lands.

JOHN S. SHONKWILER, Ph.D. Missouri (Columbia) 1979. A semi-annual econometric model of the *U.S.* feed grain-livestock economy.

JOHN W. SIEBERT, Ph.D. California (Berkeley) 1979 Almonds, bees, and externalities in the California agricultural economy.

EVERETTE B. SIGMON, Ph.D. Clemson 1978. Migration patterns and revealed values of environmental quality: An empirical extension of the Tiebout hypothesis.

MARGARET E. SLADE, Ph.D. George Washington 1979. An econometric model of the domestic copper and aluminum industries: The effects of higher energy prices and declining ore quality on metal substitution and recycling

J BARRY SMITH, Ph.D Western Ontario 1978. An analysis of optimal replenishable resource management under uncertainty.

ROLAND D. SMITH, Ph.D. Purdue 1978. An application of perceptual mapping as an approach to evaluate new product concepts for new product development

SEYOUM SOLOMON, Ph.D. Ohio State 1978. An economic analysis of alternative grain distribution systems in western Ohio.

DANIEL F. SPULBER, Ph.D. Northwestern 1979. Studies in adaptive investment planning. Research and development, rolling plans, and renewable resources.

WILLIAM S. STEWART, Ph.D. Mississippi 1977. An economic analysis of agricultural lime vending: A case study.

JAMALLUDIN B. SULAIMAN, Ph.D. Ohio State 1978. The role of industrial sites in increasing the industrial development potential of Appalachian Ohio.

DAMRONGSAK P. TASANASANTA, Ph.D. Washington State 1979. Sources of agricultural output growth in Thailand, 1950–76.

MARK A. THAYER, Ph.D. New Mexico 1978. Valuation of environmental damages.

CYNTHIA P. TINBERG, Ph.D. Michigan State 1978. An economic analysis of agricultural land-use policies for southern Michigan.

BOONJIT TITAPIWATANKUN, Ph.D. Minnesota 1979. Analysis of export demand for Thai tapioca.

FRANCIS CHIH-HUANG TUAN, Ph.D. Illinois 1979. The economic impact of coal strip mining on rural agricul

ture and communities: The case of Fulton and Knox Counties.

TIMOTHY J. TYRELL, Ph.D. Cornell 1979. An application of the multinomial logit model to predicting the pattern of food and other household expenditures in the northeastern United States.

RANDALL WESTGREN, Ph.D. Purdue 1979. The economics of corn quality measurement and grading in the wet milling industry.

MICHAEL E. WETZSTEIN, Ph.D. California (Davis) 1978 An econometric analysis of California wilderness use with emphasis on measuring the effects of introducing new areas.

RICK C. WHITACRE, Ph.D. Illinois 1979. An evaluation of Japanese agricultural trade policies with a multi-region-multicommodity model

MICHAEL W. WOOLVERTON, Ph D. Missouri (Columbia) 1978 Analysis of grain barging rates and costs on the Mississippi River

JOHN F. YANAGIDA, Ph D. Illinois 1978. Productivity analysis on a sample of Illinois grain farms, 1972 76

CHIN W. YANG, Ph D. West Virginia A critical analysis of spatial commodity modeling: The case of coal

JOSEPH YASSOUR, D.B.A Harvard 1979 Decision support systems: The case of price policy for rice in the Philippines

RICHARD D. YOUNG, Ph D California (Santa Barbara) 1979 Some aspects of production from an ocean fishery.

DAVID ZILBERMAN, Ph.D California (Berkeley) 1979 A putty clay approach to environmental quality control.

CHARLES F ZIMMERMANN, Ph D. Cornell 1978 Uranium resources on federal lands: An evaluation of policy alternatives

Manpower, Labor, and Population; including Trade Unions and Collective Bargaining

SAEED ABEDIN-DORKOOSH, Ph D Southern California 1979. Destination choice of interstate migrants to selected areas in California.

KATHRYN H. ANDERSON, Ph.D. North Carolina State 1978. The effect of child mortality on the quantity and quality of children in Guatemala

BARBARA K ATROSTIC, Ph D Pennsylvania 1978 The importance of nonpecuniary job characteristics in the labor supply decision.

RONALD M. AYERS, Ph.D Tulane 1978. Male-female earnings differentials: Theoretical and empirical analysis

JEFFREY D. BAKER, Ph D North Carolina (Chapel Hill) 1979 Specific and general training: A study of Air Force Academy graduate promotions and attrition.

GERARD BELANGER, Ph.D. Western Ontario 1978. Structural and locational determinants of regional employment growth: A case study of the Canadian pulp and paper industry.

PAUL B. BENNETT, Ph.D. Princeton 1979. Job quits and unemployment among young workers.

NANCY BIRDSALL, Ph.D. Yale 1979. Siblings and schooling in urban Colombia.

DANIEL W. BOOTHBY, Ph.D. California (Berkeley) 1978. The occupational segregation of women.

WILLIAM R. BOWMAN, Ph.D Maryland 1978. Human capital model of MDTA post-training earnings.

BERNARD BRODY, Ph.D. McGill 1979. The Canadian government geographic public service wage policy and the letter carrier case (1972-73).

GARY T. BURTLESS, Ph.D. Massachusetts Institute of Technology 1977. Taxes, transfers, and preferences for work among married black men: The Gary income maintenance experiment

TIMOTHY CARR, Ph D. Ohio State 1978. A comparative study of the duration of unemployment of young and middle-aged men.

TARSICIO CASTANEDA CAMACHO, Ph.D. Chicago 1979 Fertility, child schooling, and the labor force participation of mothers in Colombia, 1977.

CAROL M CLARK, Ph.D. Michigan 1979. Relation of economic and demographic factors to household decisions regarding education of children in Guatemala.

KIM B. CLARK, Ph.D Harvard 1978. Unions and productivity in the cement industry.

KATHLEEN P. CLASSEN, Ph D. California (Los Angeles) 1978 Unemployment insurance, the duration of unemployment, and post-unemployment earnings.

HENRY J COLEMAN, JR., Ph.D. California (Berkeley) 1978 Perceived environmental uncertainty, managerial strategy, and subunit power relationships.

JEAN M CORBO, Ph.D Boston College 1979. The *U.S.* tax structure and its effect on the allocation of time among work, leisure, and housework.

WILLIAM CURINGTON, Ph.D. Syracuse 1979. The impact of OSHA on the frequency and severity of occupational injuries

LINDA P. DATCHER, Ph.D Massachusetts Institute of Technology 1978 The effect of higher women's labor force participation rates on the relative earnings of black and white families.

ALAN E DILLINGHAM, Ph.D. Cornell 1979. The injury risk structure of occupations and wages.

DAVID A DILTS, Ph.D Indiana 1978. The impact of collective bargaining on the efficient delivery of health care services.

WENDY K DOBSON, Ph.D. Princeton 1979. The economic determinants of rising age at marriage: A case study of South Korea.

MARIE-PAULE J DONSIMONI, Ph.D. Harvard 1978. An analysis of trade union power: Structure and conduct of the American labor movement.

STUART B DORSEY, Ph.D. Washington (St Louis) 1978. The occupational licensing queue.

ALI M. EL-HUNI, Ph.D. Oklahoma State 1978. Determinants of labor force participation of women: The case of Libya.

MARIANNE V. FELTON, Ph.D. Indiana 1979. The economics of the creative arts: The case of the composer.

RANDALL K. FILER, Ph.D. Princeton 1979. The influence of affective traits and cognitive skills on the wage equation.

LAWRENCE FOGLI, Ph.D. California (Berkeley) 1978. Job analysis and organizational effectiveness.

BERNARD F. FORTIN, Ph.D. California (Berkeley) 1979. Labor supply analysis of low-income families in the Province of Quebec.

AUGUSTINE K. FOSU, Ph.D. Northwestern 1979. Choice of fringe benefits as a form of labor earnings: A theoretical and empirical analysis

CHARLES E. FRALEY, Ph.D. Oklahoma 1978. The impact of state aid on the funding of public education in Oklahoma: Incentive and equalization.

JAMES G FRANK, Ph.D. Queen's (Kingston) 1979 Labor supply estimates: The case of single individuals in Canada.

ROGER S. FRANTZ, Ph.D Washington State 1978. Beyond allocative efficiency· The role of psychological factors in worker motivation, career choice, and industrial mobility.

EUGENE B. GENDEL, Ph.D. Boston 1979 The impact on salaries of local occupation-specific labor market imbalances. The case of public school teachers in the Boston SMSA, 1960–75.

SANDRA GLEASON, Ph.D. Michigan State 1978. A benefit-cost analysis of institutional training programs in Michigan prisons.

CHRISTINE A. GREENHALGH, Ph.D. Princeton 1978 Labor supply and earnings in Great Britain with special reference to married women.

BRUCE C. GREENWALD, Ph.D. Massachusetts Institute of Technology 1978. Adverse selection in the labor market.

DAVID GUZICK, Ph.D. New York 1979 A monopolistic competition model of physician services markets.

TERRENCE HALPIN, Ph.D. Michigan State 1978. Three essays on the effects of experience rating in UI.

RICHARD B. HANSEN, Ph.D. Nebraska (Lincoln) 1979. Determinants of faculty salaries in selected departments at the University of Nebraska-Lincoln.

BENJAMIN HAWKINS, Ph.D. Kentucky 1979. The role of strike activity in public employment.

BRIAN P. HESHIZER, Ph.D. Wisconsin (Madison) 1978. Employee benefits, trade unions, and the union impact of employee benefits in manufacturing.

TERUYUKI HIGA, Ph.D. Illinois 1978. An analysis of the economic effects of the Food Stamp Program upon consumption and work effort.

LEON HOKE, Ph.D. Pittsburgh 1979. Personal and family determinations of work experience of youth in the Pittsburgh SMSA.

JANET C. HUNT, Ph.D. South Carolina 1978. Intrafamily allocation of time to nonmarket production.

PETER JACKSON, Ph.D. Harvard 1979. Wages and employment: An evaluation of the Massachusetts Balance of State Ceta Program, 1974–76.

MICHAEL S. JOHNSON, Ph.D North Carolina (Chapel Hill) 1978. The intraurban migration process and local government activity.

BRUCE E KAUFMAN, Ph.D. Wisconsin (Madison) 1978 Wage-price expectations and cyclical strike activity.

ROGER T KAUFMAN, Ph.D. Massachusetts Institute of Technology 1978. An international comparison of unemployment rates. The effects of job security and job continuity.

THIENCHAY KIRANANDANA, Ph.D Duke 1978 An economic study of fertility determination among rural and urban Thais.

ANSFRID L KOLBE, Ph D Massachusetts Institute of Technology 1979 An investigation of efficient mixes of personnel policies to maintain physicians.

CHRISTEL KONANG, Ph.D. Southern California 1978. Middle-aged men in the labor force: A graphical and empirical exploration of health capital, disability insurance, and time allocated to work

SUNG YEAL KOO, Ph.D Hawaii 1979 A study of fertility and labor force participation of married women in Korea

KENT KUNZE, Ph.D. Utah 1979 The effects of age composition and change in vital rates on nineteenth-century population estimates from new data.

LEONARD P. LARDARO, Ph.D Indiana 1979 The effect of unemployment insurance on the job search behavior of young adults in Indiana.

FRED LAZAR, Ph.D. Harvard 1978. An analysis of male racial unemployment rate differentials.

NEVILLE LEE, Ph D. Pittsburgh 1978. The influence of local government expenditures on migration patterns New wine in an old bottle.

EVELYN L. LEHRER, Ph.D. Northwestern 1978. Women's allocation of time over the life cycle: An econometric study.

JOHN P. LEIGH, Ph.D. Wisconsin (Madison) 1979. Earnings, risk, and the wage structure of industry.

LINDA S. LEIGHTON, Ph.D. Columbia 1978. Unemployment over the work history: Structure, determinants, and consequences.

STUART A. LOW, Ph.D. Illinois 1979. On-the-job and full-time job search.

DAVID M. R LUSKIN, Ph.D. Rochester 1979. The economics of minimum wage laws.

ROBERT B. MCBRIDE, Ph.D. Wisconsin (Madison) 1978. The economics of heroin distribution in the United States.

GLENN M. MACDONALD, Ph.D. Rochester 1979. Person-specific information in the labor market

JOHN M. MCDOWELL, Ph.D. California (Los Angeles) 1979. Academic productivity profiles, career choice: Academic women.

RONALD T MCHUGH, Ph D Pennsylvania State 1979 The microeconomics of water supply and demand in the northeastern United States

SHOLEH MAANI-ENTESSARI, Ph.D Illinois 1978 Study of female labor participation and fertility: A cross-cultural approach

THOMAS MACURDY, Ph D Chicago 1978 Econometric model of the labor supply in a life setting

DOUGLAS W MARDFIN, Ph D Hawaii 1979 The economic demand for abortion

JOHN C. MARTIN, Ph D George Washington 1978 The labor market of the U.S. shipbuilding industry

DANIEL J. MECKSTROTH, Ph D Cincinnati 1978 Relative wage effect in the contract construction industry

SUSAN F MEIVES, Ph D Wisconsin (Madison) 1978 Part-time work A multiperspective analysis

FARHAD MIRHADY, Ph D. Southern California 1978 A theoretical model of the optimal distribution of labor derived income

OLIVIA A MITCHELL, Ph D Wisconsin (Madison) 1978 The postwar labor supply of nonmarried and married women

PETER J MONTIEL, Ph D Massachusetts Institute of Technology 1978 Unemployment and structural transformation in Puerto Rico

SYLVIA F MOORE, Ph D Ohio State 1978. The short-run effects of marital disruption on the labor supply behavior of young women

PHILIP I MOSS, Ph D Massachusetts Institute of Technology 1978 Manpower programs and criminal justice reform: A case study of pretrial diversion

WILLIAM S MOUNTS, Ph D Georgia 1977 An analysis of the full-time/part-time work choice and wages of married women.

MARYANNE M MOWEN, Ph D Arizona State 1979 The reservation wage and unemployment duration A theoretical and empirical investigation

CHARLES F. MUELLER, Ph D Boston College 1979 Labor force migration. A behavioral approach

CRAIG A OLSON, Ph D. Wisconsin (Madison) 1978 Trade unions, wages, occupational injuries: An empirical analysis

ARTHUR H. PADILLA, Ph.D North Carolina (Chapel Hill) 1978 Race in economics: A human capital inquiry.

FUNKOO PARK, Ph.D Minnesota 1978. Analysis of labor earnings in South Korea, 1973. A human capital approach.

TIMOTHY J PERRI, Ph.D. Ohio State 1978. The health of young men: Implications for schooling and earnings

MATHANA PHANANIRAMAI, Ph D Hawaii 1979. A micro-economic analysis of the determinants of fertility in Thailand

MARY K PLANTES, Ph.D. Massachusetts Institute of Technology 1979. Work experience, economic activity, and lifetime earnings: An intercohort analysis.

NGUYEN QUAN, Ph D Michigan State 1978. Union effects on the size distribution of earnings.

CHARLES A RAMBECK, Ph D. Minnesota 1979. Family investments in migration.

CYNTHIA C RENCE, Ph D California (Berkeley) 1978. The effects of labor market mobility on earnings progression for women.

JAMES R ROOK, Ph D North Carolina State 1979. Labor skill composition, elasticities of substitution, and industrial migration: The case of cotton textiles.

IGNACIO SANTILLANA DEL BARRIO, Ph.D Indiana 1978 The economic determinants of internal migration: A case study of Spain, 1960 73

DAVID D SCHRECK, Ph D British Columbia 1978. Occupational segregation by sex

ROBERT L SELF III, Ph D North Carolina State 1978 The demand and supply of dental services.

LINDA J SHAFFER, Ph D Northwestern 1978 Job mobility and labor market segmentation

MICHAEL SHALEV, Ph D. Wisconsin (Madison) 1979 The strike Theoretical and empirical studies of industrial conflict across time and nations

FRANCIS SHIEH, Ph D California National Open University 1978 Work and study cycle theory Applications for optimum human development through open university education.

WILLIAM A SHIPMAN, JR., Ph D Pittsburgh 1978 An econometric study of the teenage labor market and school enrollment behavior

MICHAEL E SIMMONS, Ph D Washington State 1979 An inquiry into the labor market behavior of black youth A cross-sectional analysis of the National Longitudinal Survey data for 1968 and 1971

ALLEN M SINGER, Ph D George Washington 1979 Federal policy and the labor market for bioscientists.

JAMES G SPENCE, Ph.D. Oklahoma 1979. An analysis of the impact of the Ceta, Title I Institutional Skill Training Program in Pinellas County, Florida.

JAN SVEJNAR, Ph D. Princeton 1979. Economics of co-determination and trade unions Theory and empirical evidence from Germany

MARGARET J B SWEENEY, Ph.D Northwestern 1979. Family planning in a stochastic and dynamic world.

CAROL A SZERSZEN, Ph.D Illinois 1979. The impact of regulation on collective bargaining in the commercial airline industry

OMO A TAIGA, Ph D Temple 1978. The impact of skill training on high school dropouts: An econometric analysis

THOMAS N TAYLOR, Ph.D. North Carolina State 1978.

An economic analysis of the spatial distribution of veterinarians in private practice.

LUIS TOHARIA, Ph.D. Massachusetts Institute of Technology 1979. The division of labor and the historical development of the internal labor market: A case study of the McCormick Works of Chicago, 1848–1902.

SHUI-LIANG TUNG, Ph.D. Hawaii 1979. An econometric simulation of the effects of fertility control on the economy of Taiwan.

JACOB W. ULVILA, D.B.A. Harvard 1979. Decisions with multiple objectives in integrative bargaining.

CHARLES J. UNTIET, Ph.D. Stanford 1978. Hours of search and the reservation wage: A study of the job search decisions of low-income, nonworking women.

FRANCOIS VAILLANCOURT, Ph.D. Queen's (Kingston) 1978. A study of the determinants of the labor earnings of males in Quebec in 1970

JENNIFER L WARLICK, Ph D. Wisconsin (Madison) 1979. An empirical analysis of participation in the Supplemental Security Income Program among aged eligible persons

MICHAEL M. WEINSTEIN, Ph D Massachusetts Institute of Technology 1979 Some redistributive and macroeconomic impacts of the National Industrial Recovery Act, 1933-35

STEPHEN WERNER, Ph.D. State University of New York (Albany) 1978. Improving the work habits of supported workers

EVERETTE E. WHITE, Ph.D Virginia Polytechnic Institute 1979. An economic analysis of the causes of unionization of college faculty

JACK E. WHITING, Ph D. Cornell 1979 Compensating wage differentials and pension coverage The implicit market for pensions.

KATHRYN E YAEGER, Ph.D Princeton 1978. Modal choice in the demand for child care by working women A multinomial logit analysis with quality adjustment.

Welfare Programs; Consumer Economics; Urban and Regional Economics

REINALDO I ADAMS, Ph.D. Ohio State 1979. Agricultural adjustments to Brazil's alcohol program—A regional economic analysis.

JAMES R. AHRENHOLZ, Ph.D Missouri (Columbia) 1979. The effects of coal-related development on communities in the Mendota region.

JOSEPH D. ALBERT, Ph.D. Georgia State 1978 The effect of increased energy prices on urban spatial structure

JAMES L. ALTMANN, Ph.D. Wisconsin (Milwaukee) 1978. Analysis, extension, and application of the Mills-Muth urban land use model.

NABIL S AMMARI, Ph.D. Southern California 1979. Theoretical hypotheses underlying the gravity law of spatial interaction evaluated with appropriate statistical tests

CHARLES R. ANTHONY, Ph.D. Pennsylvania 1979. Health, population, and income in rural Nepal.

LOUIS G. ASCAH, Ph.D. McGill 1979. Government and private pensions in Canada.

ROBERT ATEN, Ph.D. New York 1978. A calculation of the excess burden resulting from a rental payment subsidy when compared with a property tax exemption to low-income housing in New York City.

HADI-MOHAMED BARAZI, Ph.D. State University of New York (Binghamton) 1979. Further investigation into the role of education in the demand for health care.

RALPH M. BRAID, Ph.D. Massachusetts Institute of Technology 1979. Theoretical models of urban housing as durable capital.

BARBARA BROWN, Ph.D. City (New York) 1979. Demand microeconomics of urban residential location.

SAM BUCOVETSKY, Ph.D. Yale 1979. Housing markets and the local public sector in a metropolitan area: A theoretical analysis.

SEEWOONUNDUN BUNJUN, Ph.D. Pennsylvania State 1979 Quality of life indices and their role in explaining migration.

JAMES D. BURNELL, Ph D. Illinois 1978. Regional resource development in an open economy.

RICHARD V BUTLER, Ph.D. Massachusetts Institute of Technology 1977. Hedonic indexes of urban housing: Theory and problems of cross-sectional estimation.

HENRY C. CAIN, Ph D. Florida State 1978 Redistribution effects of funding Florida's elementary and secondary educational system.

GLENN B. CANNER, Ph.D. Brown 1979. Redlining and mortgage lending practices.

MARY A. CASSELLA, Ph.D. Boston 1979. A model of the household's auto usage and ownership behavior.

CHING-HUEI CHANG, Ph D. Pennsylvania State 1978 An economic analysis of occupational disease A theoretical and empirical investigation.

CYRIL F. CHANG, Ph.D. Virginia 1979. The financing of public schools in Virginia: An econometric analysis.

KWAI-FONG CHUI, Ph.D. California (San Diego) 1979 Cost of medical care: Quality vs. time.

ALAN L. COHEN, Ph.D. Wisconsin (Madison) 1979. Economics, marital instability, and race.

GEORGE A. COLLIER, Ph D. Oklahoma State 1978. The growth center strategy in the Southern Oklahoma Development Association's development.

DOUGLAS A. CONRAD, Ph.D. Chicago 1978. The incentive effects of medical malpractice claims: Theory and empirical tests.

PAUL R. CULLINAN, Ph.D. Syracuse 1979. The effect of health on the early retirement decision.

WILLIAM F. DAVIDGE, Ph.D. Mississippi 1976. The attractive influence and current adequacy of Mississippi industrial location factors, 1963–72.

CHARLES F. DEKAY, Ph.D. Johns Hopkins 1979. The demand for cable television in urban markets.

MICHAEL K. DUFFY, Ph.D. Chicago 1979. Interurban migration: Household residence relocation and empirical analysis.

BRUCE H. DUNSON, Ph.D. Harvard 1979. A model of search behavior in rental housing markets: An empirical study.

STANLEY F. DUOBINIS, Ph.D. Ohio State 1979. An econometric model of the Chicago SMSA.

WILLIAM EMPEY, Ph.D. Boston College 1979. Regional investment in Canada.

JAMES C. FAWCETT, Ph.D. California Western 1978. SANDCAST: An econometric forecasting model of the San Diego, California regional economy.

STEVEN A. FINKLER, Ph.D. Stanford 1978. Nonprofit hospital product-line cost assessment and interfirm allocation of production

PETER S FISHER, Ph D. Wisconsin (Madison) 1978. Circuit breaker property tax relief programs: An analysis of their rationale and design

GARRY A. FLEMING, Ph.D. Kentucky 1979 Measuring regional economic stability: An input-output approach

RICHARD W FOSTER, Ph.D. Chicago 1979 Drugs and the changing politics of public mental hospitals

JAMES FRANCIS, Ph.D. Florida State 1978 An economic analysis of vocational-technological education in Florida Returns and redistribution

WILLIAM T GANLEY, Ph D California (Berkeley) 1979 Human capital and educational reform: The development of the economics of education in the history of economic theory

DAVID A. GARVIN, Ph D Massachusetts Institute of Technology 1979 An economic analysis of university behavior

GERALD L GLANDON, Ph D. Washington 1979 The production of housing services

ARTHUR H GOLDSMITH, Ph.D Illinois 1979. Health and education interrelationships over the life cycle

BAHJAT HAFEZ, Ph.D. McMaster 1979 The design and estimation of regional economic accounts in Ontario

JOHN W. HALVERSON, Ph D. Michigan 1979. The private and social net present value of medical and dental education at the University of Michigan and analysis for policy.

MARCIA HALVORSEN, Ph D. Georgia State 1978 Black-white earnings differentials in a rural southern labor market.

DONALD HAURIN, Ph.D. Chicago 1978. Property taxation in an urban economy.

STEVEN C. HAYWORTH, Ph.D. Massachusetts Institute of Technology 1977. Institutional and analytical aspects of the origins, implementation, and selected effects of rent control in Massachusetts.

JOHN T. HEFFERNAN, Ph.D. Clark 1979. An investigation into the relationship between student achievement in an introductory statistics course employing a simulated SMSA data base and selected student characteristics.

ALLAN HOCKSTEIN, Ph.D. McGill 1979. The reallocation of long-stay hospital patients: The impact of health cost.

JAMES HODGE, Ph.D. Chicago 1979. A model of firms' investment decisions across regions.

RICHARD R. HOOVEN, Ph.D. Wisconsin (Milwaukee) 1978. The monopoly model of discrimination.

MATLUB HUSSAIN, Ph.D. McGill 1978. Interregional return migration in Canada, 1950–75.

EUGENE B. HUTCHINSON, Ph.D. Virginia 1979. An empirical analysis of multifamily housing production under Sections 221 and 236 of the National Housing Act.

TIMOTHY C. IRELAND, Ph.D. Oklahoma State 1978. An econometric model for the state of Oklahoma.

IAN J. IRVINE, Ph.D Western Ontario 1978. An analysis of the distribution of lifetime purchasing power in Canada.

EUGENE J. JENNINGS, Ph.D. Southern Methodist 1979. The demand for housing emphasizing the probability of purchase. A cross-section study.

JOHN JUREWITZ, Ph.D. Wisconsin (Madison) 1978. The internalization of environmental costs in the private electric utility industry: A theoretical analysis of alternative environmental policies under rate-of-return regulation.

LEONA A. KOCHER, Ph.D. California (Davis) 1979. Economic analysis of the effects of the proposed flammability standard for upholstered furniture.

SUSAN W KRAMER, Ph.D. Duke 1978. The redistributive effects on higher education in the United States.

RONALD N. LAFFERTY, Ph.D. California (Santa Barbara) 1978. Neighborhood land use and housing values: Hedonic price indexes of single-family homes in eastern Massachusetts with special reference to the spatial effects of nonresidential land use.

THOMAS J. LAREAU, Ph.D. Indiana 1978. A welfare analysis of alternative air pollution control policies.

MICHAEL J. LEA, Ph.D North Carolina (Chapel Hill) 1978. A simultaneous equations model of residential property value and local public expenditure determination

ROBERT H. LEE, Ph.D. Johns Hopkins 1979. Location, quality, and information in medical residency selection.

PAUL LEVY, Ph.D Chicago 1979. Agglomerate economics in a spatial setting with special emphasis on shopping centers.

PETER LEWIN, Ph.D. Chicago 1979. The economics of apartheid.

DOUGLAS O. LOVE, Ph D. Illinois 1979. Influence of city size and density on prices of various goods and services.

MARY MCCORMICK, Ph.D. Columbia 1978. The

management of retrenchment: The city of New York in the 1970's.

JAMES McDERMOTT, Ph.D. State University of New York (Binghamton) 1979. The impact of aging controls on urban housing markets.

RICHARD J. MCHUGH, Ph.D. Syracuse 1979. Local and regional finance mechanisms for mass transit.

BRIAN T. MALEC, Ph.D. Syracuse 1979. State mental health care delivery systems: Historical and empirical essays.

MICHAEL L. MARLOW, Ph.D. Virginia Polytechnic Institute 1978. An assessment of the impact of the OSHA enforcement program on the occupational safety and health market.

M. SUSAN MARQUIS, Ph.D. Michigan 1978. The costs, financing, and distributional effects of a catastrophic supplement to Medicare.

ANTHONY T. NAKAZAWA, Ph.D. California (Berkeley) 1979. Consumer preferences for housing by tenure and structure type.

D. PATRICK O'DONOGHUE, Ph.D. Washington 1979. A methodology for evaluating the cost performance of Colorado hospitals

MANUEL G. OLAVE, Ph.D. California (Santa Clara) 1979. Technology specification for a family planning service program: The Costa Rican Social Security Institute case.

THOMAS M. OLSON, Ph.D. Michigan State 1978. Nonformal education delivery systems to reach limited resource farmers in Michigan.

CAYETANO W. PADERANGA, JR., Ph.D Stanford 1979. Racial preferences and housing prices: The case of San Mateo County, California.

RAYMOND B. PALMQUIST, Ph.D. Washington 1978. Theoretical and empirical issues concerning the use of property value changes in the evaluation of public projects.

CHARLES A. PIGGOTT, Ph.D. Massachusetts Institute of Technology 1979. The demand for metropolitan area manufactured products: Estimates and implications.

CHARLES F. REVIER, Ph.D. Massachusetts Institute of Technology 1978. The incidence of intermetropolitan differentials in property taxes on rental housing

PRACHARK SAKUNTALAKSNA, Ph.D. Iowa 1978. A regional test of comparative advantage: The case of the North and South regions of the United States, 1960–71.

WILLIAM I. SIMEONE, Ph.D. Boston College 1979. Analysis of the Opportunities Industrialization Center of Providence, Rhode Island, Inc.

MARILYN J. SIMON, Ph.D. Princeton 1979. Diagnoses and medical malpractice: A comparison of negligence and strict liability systems.

ROBERT D. SIMONSON, Ph.D. Nebraska (Lincoln) 1979. Efficiency and equity in Lincoln public schools: A production function study of the relationship between student learning-growth achievement and school resource utilization

NORMAN SMITH, Ph.D. Michigan State 1978. An economic analysis of the costs of instruction in the Michigan system of higher education.

CHARLES E. SOCHER, Ph.D West Virginia 1979. Market shifts and industrial relocation Effects of rising energy prices.

GENIO A. STARANCZAK, Ph.D. Queen's (Kingston) 1979 Welfare implications of Ontario's Public Housing Program.

MARJORY S THRASHER, Ph.D. Indiana 1978. The demand for higher education in the state of Michigan

HERBERT G TRAXLER, Ph.D Florida State 1978. An investigation into the major determinants of nursing home costs.

DANIEL J. VILLEGAS, Ph.D. Stanford 1979 Do the poor pay more for food?: An analysis of retail firms and the shopping behavior of consumers

GREGORY WASSALL, Ph.D. Rutgers 1979. State industrial incentive programs.

DAVID J. WEINSCHROTT, Ph.D. California (Los Angeles) 1978. Private demand for higher education· The effect of financial aid and college location on enrollment and college choice.

NORMAN E. WEISS, Ph.D. Washington (St Louis) 1978 Intrametropolitan industrial location: A heterogeneous labor approach.

WILLIAM L. WELCH, Ph.D. Michigan State 1978. The product liability system and certain no-fault proposals

ROBERTON C. WILLIAMS, JR., Ph.D. Harvard 1978 Household headship in the United States.

BERTRAM A. WOLFE, Ph.D. Maryland 1978. Optimal user charges in an urban transportation system

SUN-TIEN WU, Ph.D. Princeton 1979. The value of urban commuting time: A Taipei case study.

DOUGLAS J. YOUNGDAY, Ph.D. Wisconsin (Madison) 1978. Voluntary provision of public goods: A theory of donations.

THE AMERICAN ECONOMIC REVIEW

VOLUME LXIX

THE AMERICAN ECONOMIC ASSOCIATION

Executive Office: Nashville, Tennessee

Editorial Office: Brown University, Providence, Rhode Island

CONTENTS OF ARTICLES AND SHORTER PAPERS

CONTENTS OF PAPERS AND PROCEEDINGS

CONTRIBUTORS TO ARTICLES AND SHORTER PAPERS

CONTRIBUTORS TO PAPERS AND PROCEEDINGS

ANNOUNCEMENT

NOTICE TO ALL GRADUATE DEPARTMENTS

The December 1980 issue of the Review will carry the seventy-seventh list of doctoral dissertations in political economy in American universities and colleges. The list will specify doctoral degrees conferred during the academic year terminating June 1980. This announcement is an invitation to send us information for the preparation of the list. This announcement supersedes and replaces a letter which was sent annually from the managing editor's office.

The *Review* will publish in its December 1980 issue the names of those who will have been awarded the doctoral degree since June 1979 and the titles of their dissertations. Dissertation abstracts will no longer be published, as these are published elsewhere.

By June 30, please send us this information on 3 × 5 cards, conforming to the style shown below, one card for each individual. Please indicate by a classification number in the right-hand corner the field in which the thesis should be classified. The classification system is that used by the *Journal of Economic Literature* and printed in every issue.

JEL Classification No. ______
Name: LAST NAME IN CAPS: First Name, Initial
Institution Granting Degree: ______
Degree Conferred (Ph.D. or D.B.A.) ______ Year ______
Dissertation Title: ______

When degrees in economics are awarded under different names, such as Business Administration, Public Administration, or Industrial Relations, candidates in these fields whose training has been *primarily in economics* should be included.

All items and information should be sent to the Assistant Editor, *American Economic Review*, Box Q, Brown University, Providence, Rhode Island 02912.

The following Statement of Ownership, Management, and Circulation is provided in accordance with the requirements, as contained in 39 U.S. Code 3685. The *American Economic Review* is owned, managed, and published by the American Economic Association, a nonprofit educational organization, located at 1313 21st Avenue So., Nashville, Davidson County, Tennessee 37212. The Managing Editor is Professor George H. Borts, *American Economic Review*, Brown University, Robinson Hall, Providence, R.I. 02912. During the preceding 12 months the average number of copies printed for each issue was 26,507, the average paid circulation, 24,284; the average *free* distribution, 393; the average number of copies distributed, 24,677. Corresponding figures for the last issue before filing: 28,000 total number copies printed; 24,750 total paid circulation; 393 free copies distributed; 25,143 total distribution.